Text Analytics with SAS®

Special Collection

Foreword by
Saratendu Sethi

The correct bibliographic citation for this manual is as follows: Sethi, Saratendu. 2019. Text Analytics *with SAS®*: *Special Collection*. Cary, NC: SAS Institute Inc.

Text Analytics with SAS®: Special Collection

ISBN 978-1-64295-184-4 (PDF)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

March 2019

Table of Contents

Free SAS® e-Books: Special Collection

In this series, we have carefully curated a collection of papers that introduces and provides context to the various areas of analytics. Topics covered illustrate the power of SAS solutions that are available as tools for data analysis, highlighting a variety of commonly used techniques.

Discover more free SAS e-books!
support.sas.com/freesasebooks

sas.com/books
for additional books and resources.

sas®
THE POWER TO KNOW®

About This Book

What Does This Collection Cover?

Frequently, organizations assume that data analytics begins and ends with structured data such as a spreadsheet or database. What happens, though, if an organization wants to analyze unstructured data, such as call center notes, customer reviews or social media posts? Enter text analytics. Text analytics allows organizations to classify raw text into meaningful categories, extract needed facts from documents, and measure sentiment. Text analytics are based on statistical, linguistic and machine learning rules, and turn unstructured data into actionable insights.

SAS offers many different solutions to analyze text. The papers included in this special collection demonstrate the wide-ranging capabilities and applications of text analytics across several industries.

The following papers are excerpts from the SAS Global Users Group *Proceedings.* For more SAS Global Forum *Proceedings,* visit the online versions of the Proceedings.

More helpful resources are available at support.sas.com and sas.com/books.

We Want to Hear from You

Do you have questions about a SAS Press book that you are reading? Contact us at saspress@sas.com.

SAS Press books are written *by* SAS Users *for* SAS Users. Please visit sas.com/books to sign up to request information on how to become a SAS Press author.

We welcome your participation in the development of new books and your feedback on SAS Press books that you are using. Please visit sas.com/books to sign up to review a book

Learn about new books and exclusive discounts. Sign up for our new books mailing list today at https://support.sas.com/en/books/subscribe-books.html.

Foreword

Text analytics, also known as text analysis or text mining, is the automated process of deriving important information from unstructured text data. The study of text analytics started around the 1950s when researchers attempted to analyze human language through computational and linguistic methods. Since then, text analytics has grown into an interdisciplinary field by expanding the analysis of unstructured data to apply approaches from information theory, statistics, machine learning, and artificial intelligence. These developments, along with the exponential increase in the computational power of computers and the emergence of big data, have led organizations to foster data-positive cultures that rely on highly sophisticated applications to make business decisions based on analysis of internal documents, internet, social media, and speech data. Today, text analytics helps solve a variety of everyday business problems – such as managing and interpreting notes, assessing risk or fraud, and incorporating customer feedback for earlier problem resolution.

SAS® Text Analytics is designed for business analysts, domain experts, research analysts, linguists, knowledge workers, and data scientists who need to analyze large amounts of unstructured data to glean new insights. It offers powerful tools for consolidating, categorizing, and retrieving information across an enterprise through supervised and unsupervised machine learning, deep learning, linguistic rules, entity extraction, sentiment analysis, and topic detection.

SAS provides many different solutions to investigate and analyze text and operationalize decisioning. Several impressive papers have been written to demonstrate how to use these techniques. We have carefully selected a handful of these from recent Global Forum contributions to introduce you to the topic and let you sample what each has to offer:

Analyzing Text In-Stream and at the Edge by Simran Bagga and Saratendu Sethi

As companies increasingly use automation for operational intelligence, they are deploying machines to read, and interpret in real time, unstructured data such as news, emails, network logs, and so on. Real-time streaming analytics maximizes data value and enables organizations to act more quickly. Companies are also applying streaming analytics to provide optimal customer service at the point of interaction, improve operational efficiencies, and analyze themes of chatter about their offerings. This paper explains how you can augment real-time text analytics (such as sentiment analysis, entity extraction, content categorization, and topic detection) with in-stream analytics to derive real-time answers for innovative applications such as quant solutions at capital markets, fake-news detection at online portals, and others.

Harvesting Unstructured Data to Reduce Anti-Money Laundering (AML) Compliance Risk by Austin Cook and Beth Herron

As an anti-money laundering (AML) analyst, you face a never-ending job of staying one step ahead of nefarious actors (for example, terrorist organizations, drug cartels, and other money launderers). One area gaining traction in the financial services industry is to leverage the vast amounts of unstructured data to gain deeper insights. This paper explores the potential use cases for text analytics in AML and provides examples of entity and fact extraction and document categorization of unstructured data using SAS® Visual Text Analytics.

Invoiced: Using SAS® Text Analytics to Calculate Final Weighted Average Price by Alexandre Carvalho

SAS® Contextual Analysis brings advantages to the analysis of the millions of electronic tax notes issued in the industry and improves the validation of taxes applied. Tax calculation is one of the analytical challenges for government finance secretaries. This paper highlights two items of interest in the public sector: tax collection efficiency and the calculation of the final weighted average consumer price. SAS Contextual Analysis enables the implementation of a tax taxonomy that analyzes the contents of invoices, automatically categorizes a product, and calculates a reference value of the prices charged in the market.

Using SAS® Text Analytics to Assess International Human Trafficking Patterns by Tom Sabo and Adam Pilz

The US Department of State (DOS) and other humanitarian agencies have a vested interest in assessing and preventing human trafficking in its many forms. A subdivision within the DOS releases publicly facing Trafficking in Persons (TIP) reports for more than 200 countries annually. These reports are entirely freeform text, though there is a richness of structure hidden within the text. How can decision-makers quickly tap this information for patterns in international human trafficking? This paper showcases a strategy of applying SAS® Text Analytics to explore the TIP reports and apply new layers of structured information. Specifically, we identify common themes across the reports, use topic analysis to identify a structural similarity across reports, identifying source and destination countries involved in trafficking, and use a rule-building approach to extract these relationships from freeform text.

An Efficient Way to Deploy and Run Text Analytics Models in Hadoop by Seung Lee, Xu Yang, and Saratendu Sethi

Significant growth of the Internet has created an enormous volume of unstructured text data. In recent years, the amount of this type of data that is available for analysis has exploded. While the amount of textual data is increasing rapidly, an ability to obtain key pieces of information from such data in a fast, flexible, and efficient way is still posing challenges. This paper introduces SAS® Contextual Analysis In-Database Scoring for Hadoop, which integrates SAS® Contextual Analysis with the SAS® Embedded Process. SAS Contextual Analysis enables users to customize their text analytics models in order to realize the value of their text-based data. The SAS Embedded Process enables users to take advantage of SAS® Scoring Accelerator for Hadoop to run scoring models. By using these key SAS technologies, the overall experience of analyzing unstructured text data can be greatly improved. The paper also provides guidelines and examples on how to publish and run category, concept, and sentiment models for text analytics in Hadoop.

Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaints by Tom Sabo

The Consumer Financial Protection Bureau (CFPB) collects tens of thousands of complaints against companies each year, many of which result in the companies in question taking action, including making payouts to the individuals who filed the complaints. Given the volume of the complaints, how can an overseeing organization quantitatively assess the data for various trends, including the areas of greatest concern for consumers? In this presentation, we propose a repeatable model of text analytics techniques to the publicly available CFPB data. Specifically, we use SAS® Contextual Analysis to explore sentiment and machine learning techniques to model the natural language available in each freeform complaint against a disposition code for the complaint, primarily focusing on whether a company paid out money. This process generates a taxonomy in an automated manner. We also explore methods to structure and visualize the results, showcasing how areas of concern are made available to analysts using SAS® Visual Analytics and SAS® Visual Statistics. Finally, we discuss the applications of this methodology for overseeing government agencies and financial institutions alike.

Exploring the Art and Science of SAS® Text Analytics: Best Practices in Developing Rule-Based Models by Murali Pagolu, Christina Engelhardt, and Cheyanne Baird

Traditional analytical modeling, with roots in statistical techniques, works best on structured data. Structured data enables you to impose certain standards and formats in which to store the data values. The nuances of language, context, and subjectivity of text make it more complex to fit generalized models. Although statistical methods using supervised learning prove efficient and effective in some cases, sometimes you need a different approach. These situations are when rule-based models with Natural Language Processing capabilities can add significant value. In what context would you choose a rule-based modeling versus a statistical approach? How do you assess the tradeoffs of choosing a rule-based modeling approach with higher interpretability versus a statistical model that is black-box in nature? How can we develop rule-based models that optimize model performance without compromising accuracy? How can we design, construct, and maintain a complex rule-based model? What is a data-driven approach to rule writing? What are the common pitfalls to avoid? In this paper, we discuss all these questions based on our experiences working with SAS® Contextual Analysis and SAS® Sentiment Analysis.

We hope these selections give you a useful overview of the many tools and techniques that are available to analyze text.

Additionally, SAS offers free video tutorials on text analytics. For more information, go to https://video.sas.com/category/videos/ and enter “text analytics” in the search box.

Saratendu Sethi

Sr Director, Advanced Analytics R&D

Saratendu Sethi is Head of Artificial Intelligence and Machine Learning R&D at SAS Institute. He leads SAS’ software development and research teams for Artificial Intelligence, Machine Learning, Cognitive Computing, Deep Learning, and Text Analytics. Saratendu has extensive experience in building global R&D teams, launching new products and business strategies. Perennially fascinated by how technology enables a creative life, he is a staunch believer in transforming powerful algorithms into innovative technologies. At SAS, his teams develop machine learning, cognitive- and semantic-enriched capabilities for unstructured data and multimedia analytics. He joined SAS Institute through the acquisition of Teragram Corporation, where he was responsible for the development of natural language processing and text analytics technologies. Before joining Teragram, Saratendu held research positions at the IBM Almaden Research Center and at Boston University, specializing in computer vision, pattern recognition, and content-based search.

Paper SAS1962-2018

Analyzing Text In-Stream and at the Edge

Simran Bagga and Saratendu Sethi, SAS Institute Inc.

ABSTRACT

As companies increasingly use automation for operational intelligence, they are deploying machines to read, and interpret in real time, unstructured data such as news, emails, network logs, and so on. Real-time streaming analytics maximizes data value and enables organizations to act more quickly. For example, being able to analyze unstructured text in-stream and at the “edge” provides a competitive advantage to financial technology (fintech) companies, who use these analyses to drive algorithmic trading strategies. Companies are also applying streaming analytics to provide optimal customer service at the point of interaction, improve operational efficiencies, and analyze themes of chatter about their offerings. This paper explains how you can augment real-time text analytics (such as sentiment analysis, entity extraction, content categorization, and topic detection) with in-stream analytics to derive real-time answers for innovative applications such as quant solutions at capital markets, fake-news detection at online portals, and others.

INTRODUCTION

Text analytics is appropriate when the volume of unstructured text content can no longer be economically reviewed and analyzed manually. The output of text analytics can be applied to a variety of business use cases: detecting and tracking service or quality issues, quantifying customer feedback, assessing risk, improving operational processes, enhancing predictive models, and many more. SAS® Visual Text Analytics provides a unified and flexible framework that enables you to tackle numerous use cases by building a variety of text analytics models. A pipeline-based approach enables you to easily connect relevant nodes that you can use to generate these models.

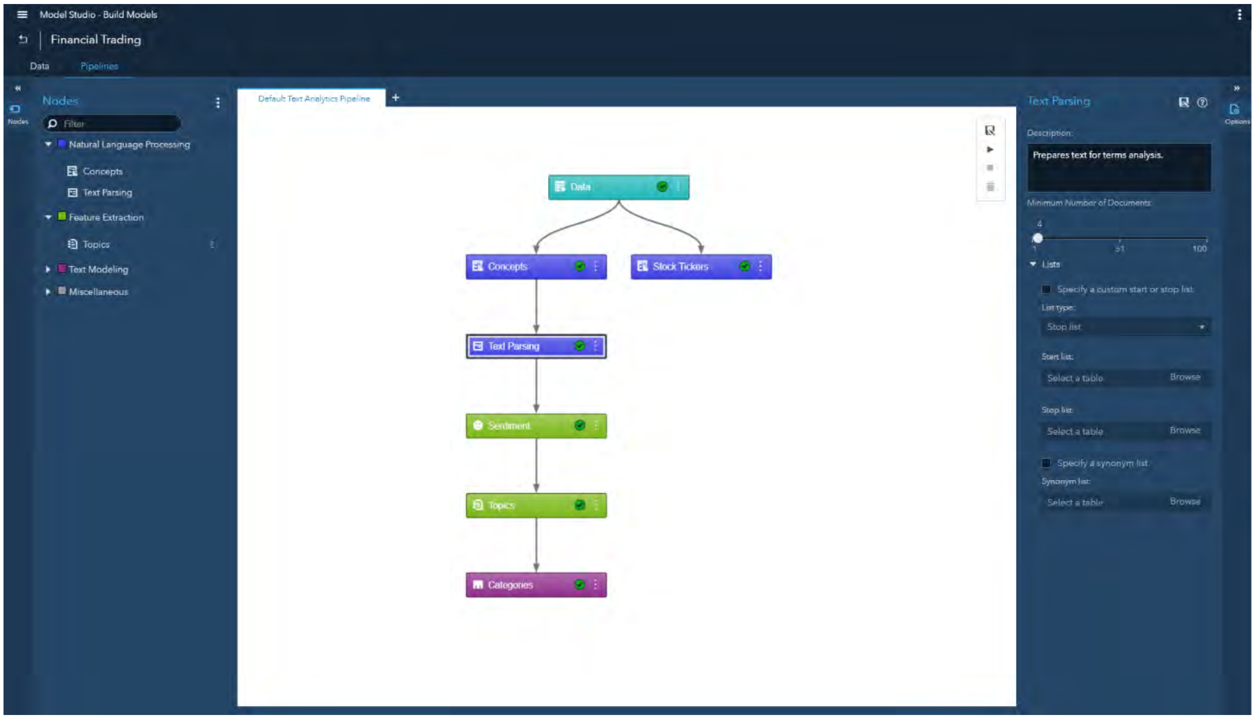
Concepts models enable you to extract entities, concepts, and facts that are relevant to the business. Topic models exploit the power of natural language processing (NLP) and machine learning to discover relevant themes from text. You can use Categories and Sentiment models to tag emotions and reveal insights and issues.

Growing numbers of devices and dependency on Internet of Things (IoT) are causing an increasing need for faster processing, cloud adoption, edge computing, and embedded analytics. The ability to analyze and score unstructured text in real time as events are streaming in is becoming more critical than ever. This paper outlines the use of SAS Visual Text Analytics and SAS® Event Stream Processing to demonstrate a complex event processing scenario. Text models for concept extraction, document categorization, and sentiment analysis are deployed in SAS Event Stream Processing to gain real-time insights and support decision making that is based on intelligence gathered from streaming events.

Big data typically come in dribs and drabs from various sources such as Facebook, Twitter, bank transactions, sensor reading, logs, and so on. The first section of this paper uses SAS Visual Text Analytics to analyze data from trending financial tweets. The latter half focuses on the deployment of text models within SAS Event Stream Processing to assess market impact and intelligently respond to each of the events or data streams as they come in.

EXTRACTING INTELLIGENCE FROM UNSTRUCTURED TEXT USING SAS VISUAL TEXT ANALYTICS

SAS Visual Text Analytics provides a modern, flexible, and end-to-end analytics framework for building a variety of text analytics models that address many use cases. You can exploit the power of natural language processing (NLP), machine learning, and linguistic rules within this single environment. The main focus of NLP is to extract key elements of interest, which can be terms, entities, facts, and so on. Display 1 demonstrates a custom pipeline that you might assemble for a text analytics processing flow. The Concepts node and the Text Parsing node give you the flexibility to enhance the output of NLP and customize the extraction process.



Description: Prepares text for terms analysis.

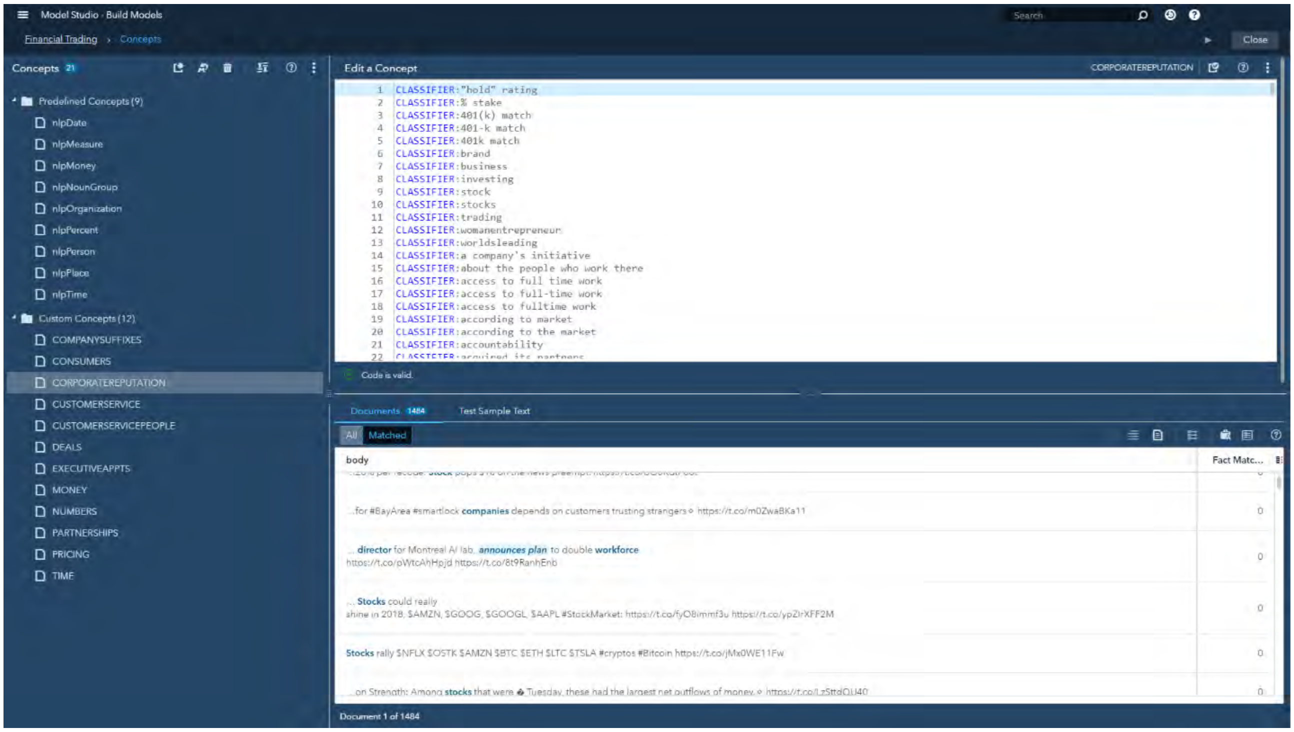
Minimum Number of Documents: 4

List type: Stop list

Display 1. Custom Pipeline in SAS Visual Text Analytics

The following list describes the role of each node in this custom pipeline.

- In the Concepts node, you include predefined concepts such as nlpDate, nlpMoney, nlpOrganization, and so on. In this node, you can also create custom concepts and extend the definitions for predefined concepts that are already built into the software. Display 2 shows some custom concepts that have been built to extract information that is related to customer service, corporate reputation, executive appointment, partnerships, and so on, and is likely to affect market trading and volatility. These custom concepts are used for associating categories to each event in SAS Event Stream Processing and will enable automatic concept extraction in future narratives.



```
CLASSIFIER:"hold" rating
CLASSIFIER:% stake
CLASSIFIER:401(k) match
CLASSIFIER:401-k match
CLASSIFIER:401k match
CLASSIFIER:brand
CLASSIFIER:business
CLASSIFIER:investing
CLASSIFIER:stock
CLASSIFIER:stocks
CLASSIFIER:trading
CLASSIFIER:womanentrepreneur
CLASSIFIER:worldsleading
CLASSIFIER:a company's initiative
CLASSIFIER:about the people who work there
CLASSIFIER:access to full time work
CLASSIFIER:access to full-time work
CLASSIFIER:access to fulltime work
CLASSIFIER:according to market
CLASSIFIER:according to the market
CLASSIFIER:accountability
CLASSIFIER:acquired its partners
```

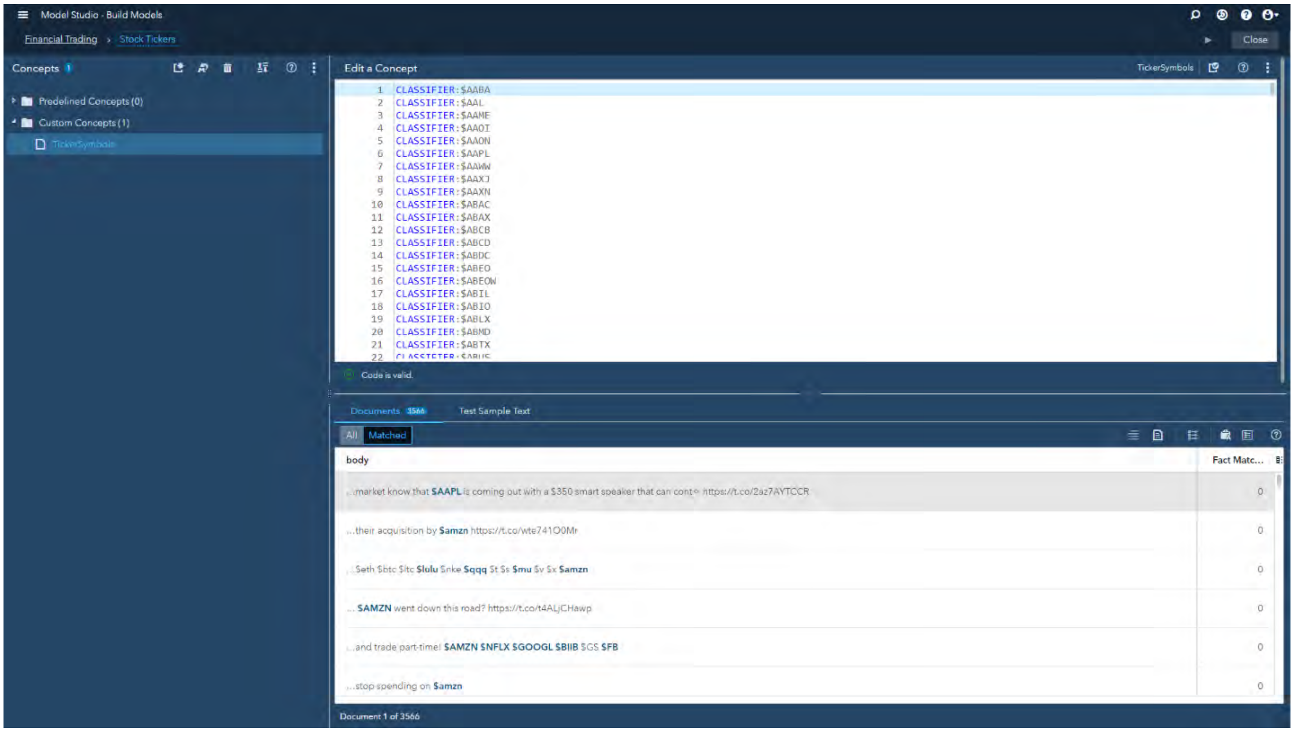
Code is valid.

body	Fact Matc...
...2018 per recode: Stock pops 3% on the news preempt: https://t.co/[illegible]	0
...for #BayArea #smartlock companies depends on customers trusting strangers ◇ https://t.co/m0ZwaBKa11	0
... director for Montreal AI lab, announces plan to double workforce https://t.co/pWtcAhHpjd https://t.co/8t9RanhEnb	0
... Stocks could really shine in 2018. $AMZN, $GOOG, $GOOGL, $AAPL #StockMarket: https://t.co/fyO8immf3u https://t.co/ypZIrXFF2M	0
Stocks rally $NFLX $OSTK $AMZN $BTC $ETH $LTC $TSLA #cryptos #Bitcoin https://t.co/jMx0WE11Fw	0
...on Strength: Among stocks that were ◆ Tuesday, these had the largest net outflows of money. ◇ https://t.co/LzSttdQU40	0

Document 1 of 1484

Display 2. Concepts Extraction in SAS Visual Text Analytics

In addition, a custom concepts model is also developed to identify stock ticker symbols in each event. This custom concept model is shown in Display 3.



```
CLASSIFIER:$AABA
CLASSIFIER:$AAL
CLASSIFIER:$AAME
CLASSIFIER:$AAOI
CLASSIFIER:$AAON
CLASSIFIER:$AAPL
CLASSIFIER:$AAWW
CLASSIFIER:$AAXJ
CLASSIFIER:$AAXN
CLASSIFIER:$ABAC
CLASSIFIER:$ABAX
CLASSIFIER:$ABCB
CLASSIFIER:$ABCD
CLASSIFIER:$ABDC
CLASSIFIER:$ABEO
CLASSIFIER:$ABEOW
CLASSIFIER:$ABIL
CLASSIFIER:$ABIO
CLASSIFIER:$ABLX
CLASSIFIER:$ABMD
CLASSIFIER:$ABTX
CLASSIFIER:$ABUS
```

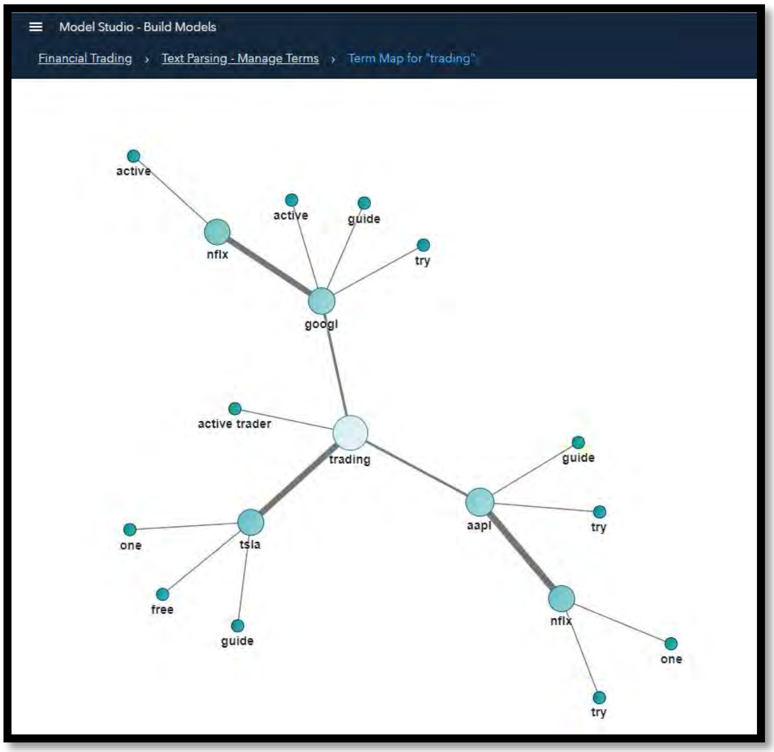
Code is valid.

body	Fact Matc...
...market know that $AAPL is coming out with a $350 smart speaker that can cont◇ https://t.co/2az7AYTCCR	0
...their acquisition by $amzn https://t.co/wte741O0Mr	0
...$eth $btc $itc $lulu $nke $qqq $t $s $mu $v $x $amzn	0
... $AMZN went down this road? https://t.co/t4ALjCHawp	0
...and trade part-time! $AMZN $NFLX $GOOGL $BIIB $GS $FB	0
...stop spending on $amzn	0

Document 1 of 3566

Display 3. Extracting Stock Ticker Symbols from Text in SAS Visual Text Analytics

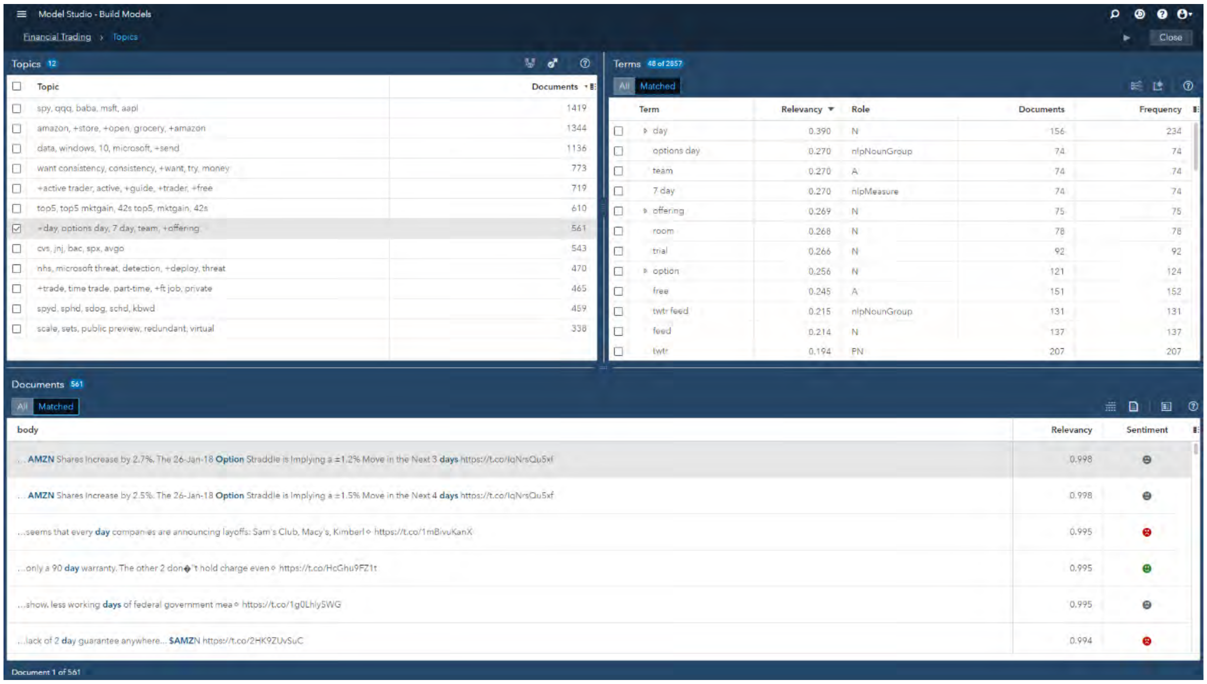
- The Text Parsing node automatically extracts terms and noun groups from text by associating different parts of speech and understanding the context. Recommended lists of Keep and Drop terms are displayed in the interactive window. After the node execution is complete, you can right-click on the node to open the interactive window and drop terms that are not relevant for downstream analysis. The Term Map within the interactive window helps you understand the association of other terms to the term "trading." See Display 4.


Model Studio - Build Models
Financial Trading › Text Parsing - Manage Terms › Term Map for "trading"
active
nflx
active
guide
try
googl
active trader
trading
guide
aapl
try
one
tsla
free
guide
nflx
one
try

Display 4. Term Map in SAS Visual Text Analytics

- The Sentiment node uses a domain-independent model that is included with SAS Visual Text Analytics. This rules-based analytic model computes sentiment relevancy for each post and classifies the emotion in unstructured text as positive, negative, or neutral. You can deploy the sentiment model in SAS Event Stream Processing to tag emotions that are associated with a post and that might affect trading decisions.

- The final list of terms from text parsing are fed into machine learning for topic detection. In the interactive window of the Text Topics node (see Display 5), you can see commonly occurring themes within a set of tweets. For example, if you select your topic of interest as "+day, options day, 7 day, team, +offering," the Documents pane shows all the tweets that mention that topic and the terms that exist within that topic, in addition to relevancy and sentiment. You can deploy the Topics model in-stream in order to capture themes as data or events are streaming in. You can also promote topics of interest into your Categories model, which you can deploy in order to classify text into multiple categories. The implementation of this example uses some categories that were created by promoting relevant topics.



☐	Topic	Documents
☐	spy, qqq, baba, msft, aapl	1419
☐	amazon, +store, +open, grocery, +amazon	1344
☐	data, windows, 10, microsoft, +send	1136
☐	want consistency, consistency, +want, try, money	773
☐	+active trader, active, +guide, +trader, +free	719
☐	top5, top5 mktgain, 42s top5, mktgain, 42s	610
☑	+day, options day, 7 day, team, +offering	561
☐	cvs, jnj, bac, spx, avgo	543
☐	nhs, microsoft threat, detection, +deploy, threat	470
☐	+trade, time trade, part-time, +ft job, private	465
☐	spyd, sphd, sdog, schd, kbwd	459
☐	scale, sets, public preview, redundant, virtual	338

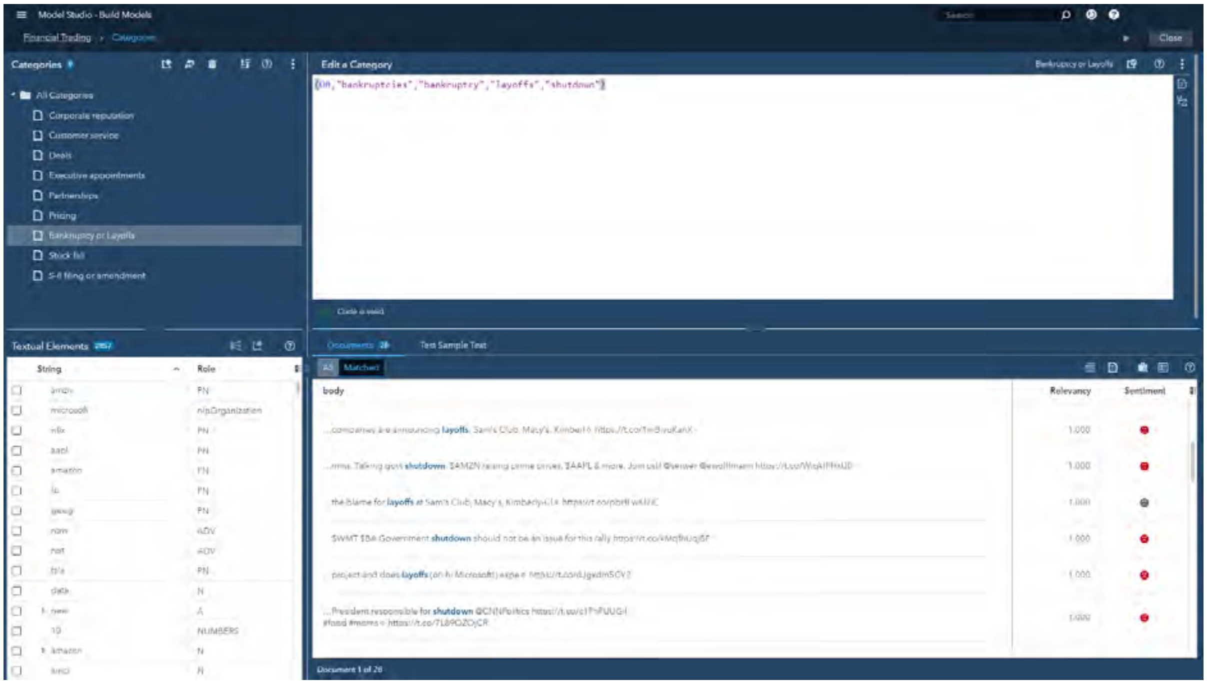
☐	Term	Relevancy ▼	Role	Documents	Frequency
☐	▸ day	0.390	N	156	234
☐	options day	0.270	nlpNounGroup	74	74
☐	team	0.270	A	74	74
☐	7 day	0.270	nlpMeasure	74	74
☐	▸ offering	0.269	N	75	75
☐	room	0.268	N	78	78
☐	trial	0.266	N	92	92
☐	▸ option	0.256	N	121	124
☐	free	0.245	A	151	152
☐	twtr feed	0.215	nlpNounGroup	131	131
☐	feed	0.214	N	137	137
☐	twtr	0.194	PN	207	207

body	Relevancy	Sentiment
... AMZN Shares Increase by 2.7%. The 26-Jan-18 Option Straddle is Implying a ±1.2% Move in the Next 3 days https://t.co/IqNrsQu5xf	0.998	●
... AMZN Shares Increase by 2.5%. The 26-Jan-18 Option Straddle is Implying a ±1.5% Move in the Next 4 days https://t.co/IqNrsQu5xf	0.998	●
...seems that every day companies are announcing layoffs: Sam's Club, Macy's, Kimberl◊ https://t.co/1mBivuKanX	0.995	●
...only a 90 day warranty. The other 2 don�'t hold charge even◊ https://t.co/HcGhu9FZ1t	0.995	●
...show, less working days of federal government mea◊ https://t.co/1g0Lhly5WG	0.995	●
...lack of 2 day guarantee anywhere... $AMZN https://t.co/2HK9ZUvSuC	0.994	●

Document 1 of 561

Display 5. Text Topics in SAS Visual Text Analytics

- In the Categories node, you see the taxonomy (Display 6) that has been designed for document categorization. You can manually extend the auto-generated rules from promoted topics and refer to the previously created concepts within your category rules. You can also use the Textual Elements table to select elements of interest that can be inserted into new rules. Multiple posts or tweets about bankruptcy or layoffs, or about an increase or decrease in the number of shares, often result in stock trading shortly thereafter. This intelligence, if available in real time, can aid in buy or sell decisions that are related to that company.



- All Categories
  - Corporate reputation
  - Customer service
  - Deals
  - Executive appointments
  - Partnerships
  - Pricing
  - Bankruptcy or Layoffs
  - Stock fall
  - S-ll filing or amendment

```
(OR,"bankruptcies","bankruptcy","layoffs","shutdown")
```

Code is valid

☐	String	Role
☐	amzn	PN
☐	microsoft	nlpOrganization
☐	nflx	PN
☐	aapl	PN
☐	amazon	PN
☐	fb	PN
☐	goog	PN
☐	now	ADV
☐	not	ADV
☐	tsla	PN
☐	data	N
☐	▸ new	A
☐	10	NUMBERS
☐	▸ amazon	N
☐	[illegible]	N

body	Relevancy	Sentiment
...companies are announcing layoffs: Sam's Club, Macy's, Kimberl◊ https://t.co/1mBivuKanX	1.000	●
...rms. Talking govt shutdown, $AMZN raising prime prices, $AAPL & more. Join us! [illegible] https://t.co/WqAlPHxUD	1.000	●
the blame for layoffs at Sam's Club, Macy's, Kimberly-Cl◊ https://t.co/pbHlwKlJJC	1.000	●
$WMT $BA Government shutdown should not be an issue for this rally https://t.co/kMqfhUqj6P	1.000	●
project and does layoffs (oh hi Microsoft) expe◊ https://t.co/iJgxdm5CV2	1.000	●
...President responsible for shutdown @CNNPolitics https://t.co/c1PhPUUGH #food #moms◊ https://t.co/7L89OZOjCR	1.000	●

Document 1 of 28

Display 6. Categorization in SAS Visual Text Analytics

SCORING FINANCIAL POSTS IN REAL TIME TO ASSESS MARKET IMPACT

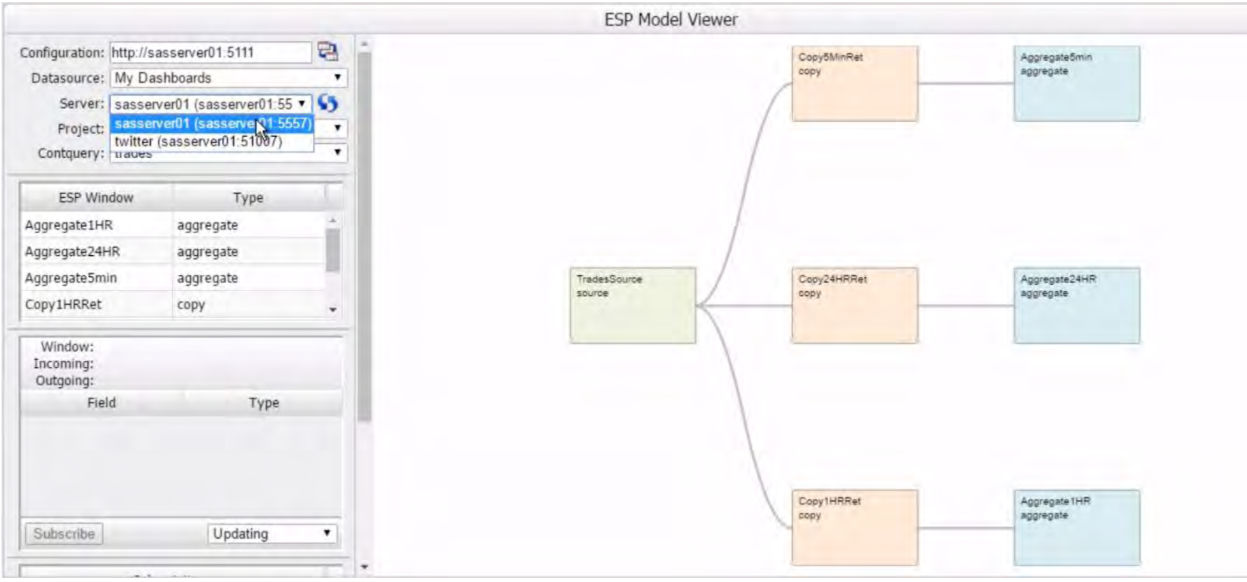
SAS Event Stream Processing is a streaming engine that enables you to analyze or score data as they stream in, rather than first storing them in the database and then analyzing and scoring them in batch. Being able to react to the clicks and events as they are coming in reduces time to action. Event stream processing can occur in three distinct places: at the edge of the network, in the stream, or on data that's at rest (out of the stream).

The SAS Event Stream Processing engine is a transformation engine that augments and adds value to incoming event streams. It is capable of processing millions of events per second. You can perform traditional data management tasks such as filtering out unimportant events, aggregating data, improving data quality, and applying other computations. You can also perform advanced analytic tasks such as pattern detection and text analysis. Events coming in from any source—sensors, Wall Street feeds, router feeds, message buses, server log files—can be read, analyzed, and written back to target applications in real time.

COMPARING STOCK TRADING WEIGHTED AVERAGE PRICE OVER THREE RETENTION PERIODS

The SAS Event Stream Processing studio is a development and testing application for event stream processing (ESP) models. An ESP model is a program or set of instructions that transforms the input event streams into meaningful output event streams. Once the models are built, they can be published into SAS Event Stream Processing for scoring.

In the ESP model presented in Display 7, the Source window (named TradesSource) is reading from one million stock trades, which are all structured data. The three Copy windows define three different levels of event retention: 5 minutes, 1 hour, and 24 hours. The three Aggregate windows create weighted average trade amounts by stock symbol.



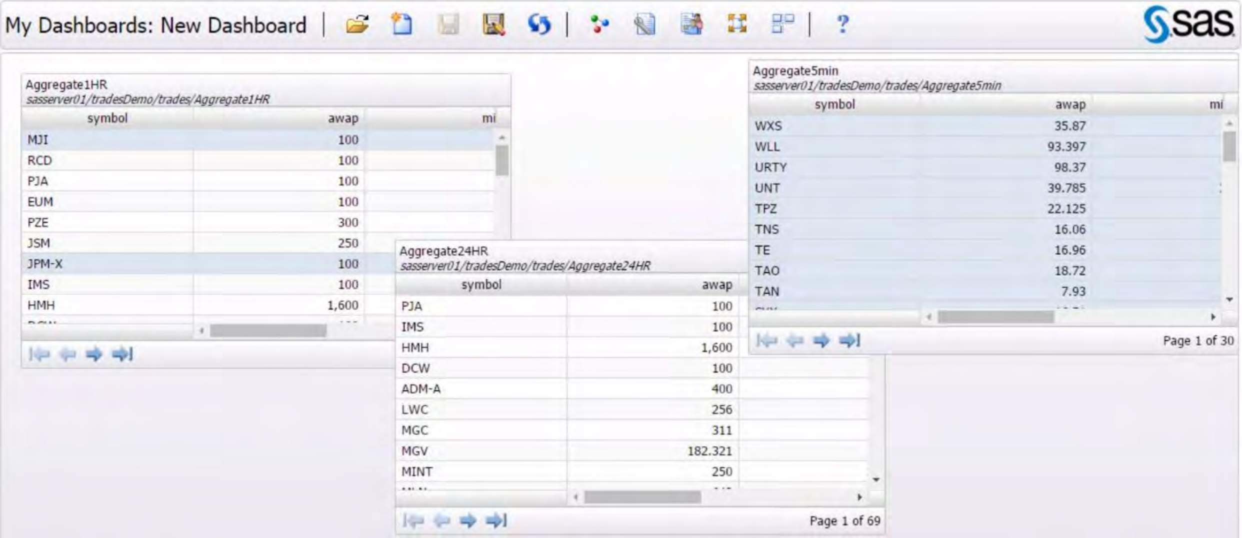
Configuration: http://sasserver01:5111
Datasource: My Dashboards
Server: sasserver01 (sasserver01:55
sasserver01 (sasserver01:5557)
twitter (sasserver01:51007)
Project:
Contquery: trades

ESP Window	Type
Aggregate1HR	aggregate
Aggregate24HR	aggregate
Aggregate5min	aggregate
Copy1HRRet	copy

Window:
Incoming:
Outgoing:

Display 7. Model Viewer in SAS Event Stream Processing

The Stream Viewer window in SAS Event Stream Processing provides a dashboard that enables you to visualize streaming events. This example creates three subscriptions for the three aggregate windows, which can viewed in the dashboard of the Stream Viewer. The dashboard in Display 8 compares the stock trading weighted average price over three retention periods: 5 minute, 1 hour, and 24 hours. The 5-minute view shows what the market is doing right now, whereas the 24-hour view shows what the full day of the market looks like.



Aggregate1HR
sasserver01/tradesDemo/trades/Aggregate1HR

symbol	awap	mi
MJI	100	
RCD	100	
PJA	100	
EUM	100	
PZE	300	
JSM	250	
JPM-X	100	
IMS	100	
HMH	1,600	

Aggregate24HR
sasserver01/tradesDemo/trades/Aggregate24HR

symbol	awap
PJA	100
IMS	100
HMH	1,600
DCW	100
ADM-A	400
LWC	256
MGC	311
MGV	182.321
MINT	250

Page 1 of 69

Aggregate5min
sasserver01/tradesDemo/trades/Aggregate5min

symbol	awap	mi
WXS	35.87	
WLL	93.397	
URTY	98.37	
UNT	39.785	
TPZ	22.125	
TNS	16.06	
TE	16.96	
TAO	18.72	
TAN	7.93	

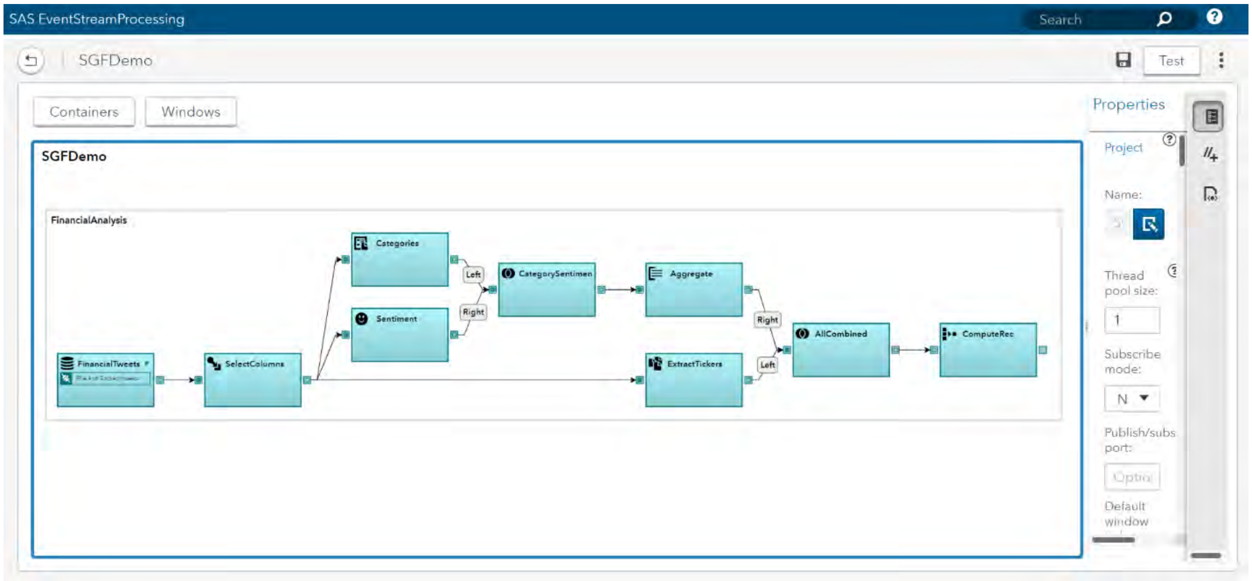
Page 1 of 30

Display 8. Dashboard Viewer in SAS Event Stream Processing

STOCK RECOMMENDATION BASED ON ANALYSIS OF UNSTRUCTURED TEXT

The models that are built using SAS Visual Text Analytics can applied in batch, in-Hadoop, in-stream, and at the edge. This section uses SAS Event Stream Processing to extract concepts, analyze sentiment about particular companies and their stock, and categorize posts as events stream in real time.

In the process defined in Display 9, tweets are continuously flowing through. The Source window (named FinancialTweets) has a retention policy of 15 minutes, which means that the analysis recommendation is based on the last 15 minutes of captured events. As the tweets come in, they are analyzed: stocks tickers are extracted, sentiment score is assigned, and the content is tagged for appropriate categories.



Name:

Thread pool size: 1

Subscribe mode: N

Publish/subscribe port: Optio

Default window

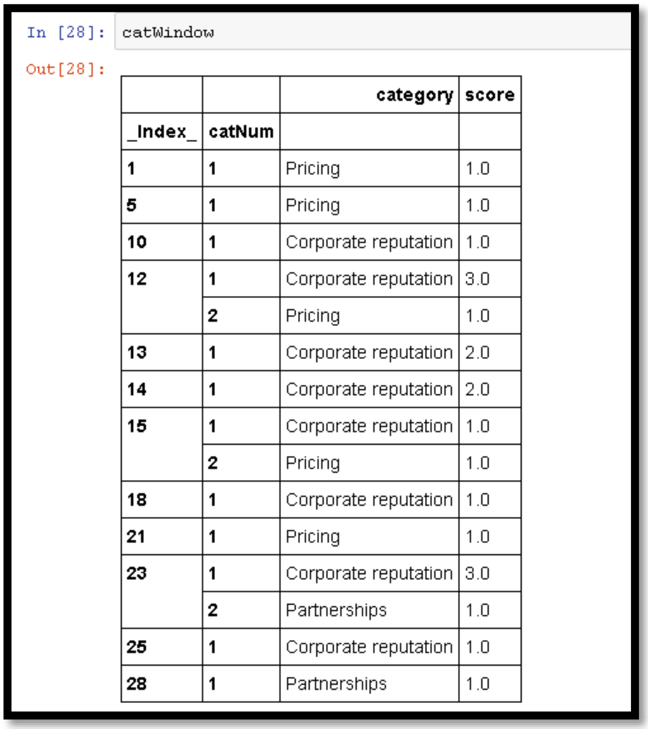
Display 9. SAS Event Stream Processing Studio

The following list describes each window in Display 9 and its role in the flow.

- FinancialTweets: This is a Source window, which is required for each continuous query. All event streams enter continuous queries by being published (injected) into a Source window. Event streams cannot be published into any other window type. Source windows are typically connected to one or more derived windows. Derived windows can detect patterns in the data, transform the data, aggregate the data, analyze the data, or perform computations based on the data. This example uses a CSV (comma-separated values) file with a small sample of tweets that are related to financial and corporate information. Because the sample is small, the results derived here are purely a proof of concept rather than a true financial analysis for all publicly traded companies. For a true streaming use case, SAS Event Stream Processing provides a Twitter adapter, which can be used to feed tweets in real time.

- SelectColumns: This Compute window enables a one-to-one transformation of input events to output events through computational manipulation of the input event stream fields. You can use the Compute window to project input fields from one event to a new event and to augment the new event with fields that result from a calculation. You can change the set of key fields within the Compute window. This example uses the SelectColumns window to filter out attributes that are not relevant for downstream analysis.

- Categories: This is a Text Category window, which categorizes a text field in incoming events. The text field can generate zero or more categories, with scores. Text Category windows are insert-only. This example uses the model file (.mco) that is generated by the **Download Score Code** option of the Categories node in SAS Visual Text Analytics. Display 10 shows the output that is generated by this window. The output lists the document ID (_Index_ column), category number (catNum column), tagged category (category column), and the relevancy score for assigned categorization (score column).



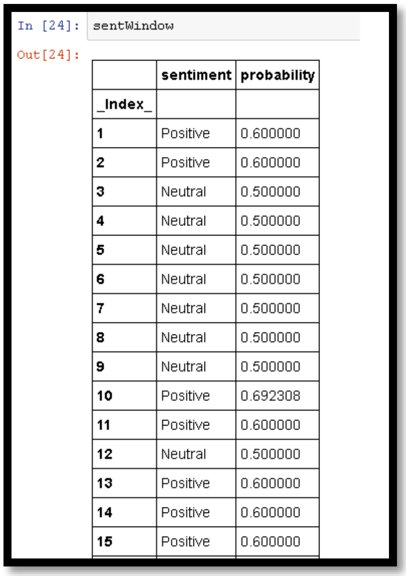
```
In [28]: catWindow
```

Out[28]:

		category	score
Index	**catNum**		
1	**1**	Pricing	1.0
5	**1**	Pricing	1.0
10	**1**	Corporate reputation	1.0
12	**1**	Corporate reputation	3.0
	2	Pricing	1.0
13	**1**	Corporate reputation	2.0
14	**1**	Corporate reputation	2.0
15	**1**	Corporate reputation	1.0
	2	Pricing	1.0
18	**1**	Corporate reputation	1.0
21	**1**	Pricing	1.0
23	**1**	Corporate reputation	3.0
	2	Partnerships	1.0
25	**1**	Corporate reputation	1.0
28	**1**	Partnerships	1.0

Display 10. Text Category Window Output

- Sentiment: This is a Text Sentiment window, which determines the sentiment of text in the specified incoming text field and the probability of its occurrence. The sentiment value is positive, neutral, or negative. The probability is a value between 0 and 1. Text Sentiment windows are insert-only. This example uses the domain-independent sentiment model file (en-base.sam), which is included in SAS Visual Text Analytics. Display 11 shows the output that is generated by this window. Upon scoring, each document in the _Index_ column is assigned an appropriate sentiment tag (in the sentiment column) along with a relevancy score (in the probability column).

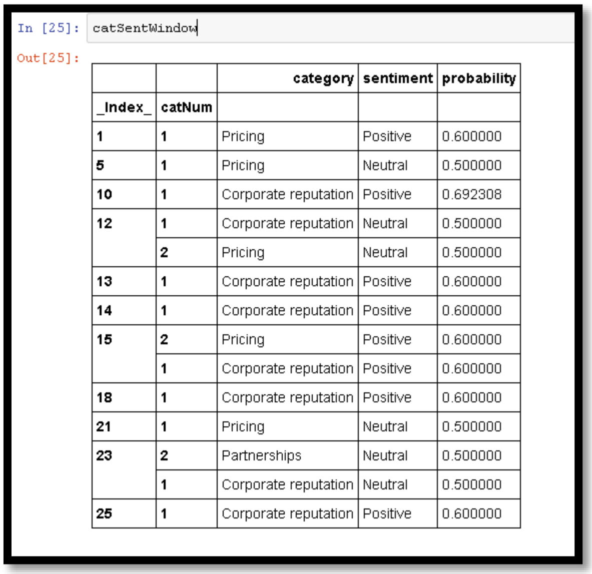


```
In [24]: sentWindow
Out[24]:
```

	sentiment	probability
Index		
1	Positive	0.600000
2	Positive	0.600000
3	Neutral	0.500000
4	Neutral	0.500000
5	Neutral	0.500000
6	Neutral	0.500000
7	Neutral	0.500000
8	Neutral	0.500000
9	Neutral	0.500000
10	Positive	0.692308
11	Positive	0.600000
12	Neutral	0.500000
13	Positive	0.600000
14	Positive	0.600000
15	Positive	0.600000

Display 11. Text Sentiment Window Output

- CategorySentiment: This is a Join window, which receives events from an input window to the left of the Join window and produces a single output stream of joined events. Joined events are created according to a user-specified join type and user-defined join conditions. This example does an inner join between the category and sentiment tables to create joined events only when one or more matching events occur on the side opposite the input event. Display 12 shows the output that is generated by the CategorySentiment window.

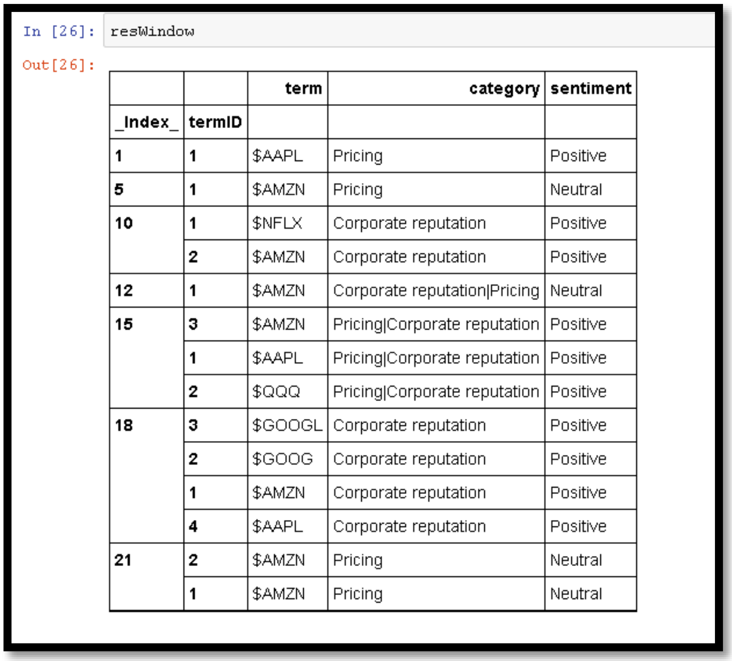


```
In [25]: catSentWindow
Out[25]:
```

		category	sentiment	probability
Index	catNum			
1	1	Pricing	Positive	0.600000
5	1	Pricing	Neutral	0.500000
10	1	Corporate reputation	Positive	0.692308
12	1	Corporate reputation	Neutral	0.500000
	2	Pricing	Neutral	0.500000
13	1	Corporate reputation	Positive	0.600000
14	1	Corporate reputation	Positive	0.600000
15	2	Pricing	Positive	0.600000
	1	Corporate reputation	Positive	0.600000
18	1	Corporate reputation	Positive	0.600000
21	1	Pricing	Neutral	0.500000
23	2	Partnerships	Neutral	0.500000
	1	Corporate reputation	Neutral	0.500000
25	1	Corporate reputation	Positive	0.600000

Display 12. Joining Category and Sentiment Output Using an Inner Join

- **Aggregate:** Aggregate windows are similar to Compute windows in that non-key fields are computed. Incoming events are placed into aggregate groups such that each event in a group has identical values for the specified key fields. This example aggregates category and sentiment information by stock ticker, as shown in Display 13.



```
In [26]: resWindow
Out[26]:
```

		term	category	sentiment
Index	termID			
1	1	$AAPL	Pricing	Positive
5	1	$AMZN	Pricing	Neutral
10	1	$NFLX	Corporate reputation	Positive
	2	$AMZN	Corporate reputation	Positive
12	1	$AMZN	Corporate reputation|Pricing	Neutral
15	3	$AMZN	Pricing|Corporate reputation	Positive
	1	$AAPL	Pricing|Corporate reputation	Positive
	2	$QQQ	Pricing|Corporate reputation	Positive
18	3	$GOOGL	Corporate reputation	Positive
	2	$GOOG	Corporate reputation	Positive
	1	$AMZN	Corporate reputation	Positive
	4	$AAPL	Corporate reputation	Positive
21	2	$AMZN	Pricing	Neutral
	1	$AMZN	Pricing	Neutral

Display 13. Joining Category and Sentiment Output with Extracted Ticker Concepts and Aggregating Categories for Each Stock Ticker

- ExtractTickers: This is a Text Context window, which is used here to call the SAS Visual Text Analytics Concepts model to extract key terms or entities of interest from text. Events that are generated from the terms can be analyzed by other window types. For example, a Pattern window could follow a Text Context window to look for tweet patterns of interest. This example combines the extracted tickers with category and sentiment information from posts.

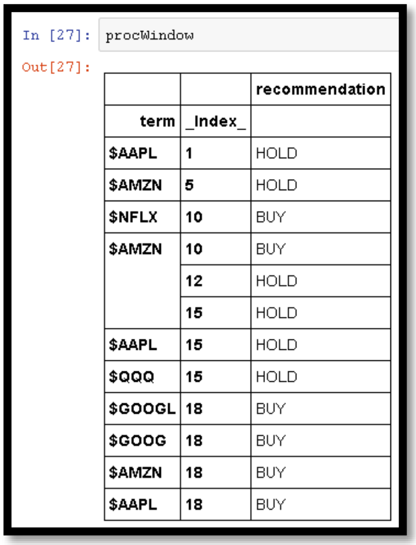
  The stock tickers are extracted by using the model file (.li) that is generated by the **Download Score Code** option of the Concepts node in SAS Visual Text Analytics. This file is also shown in Display 3.

- AllCombined: This is a second Join window; it combines output from the CategorySentiment window with output from the ExtractTickers window. Display 13 shows the output that is generated by this window. In the AllCombined output, categories and sentiment are aggregated across each stock ticker symbol within a particular document. For example, in document ID 15, $AMZN refers to both “Pricing” and “Corporate reputation” categories, with the overall sentiment being positive.

- ComputeRec: This is a Procedural window, which is a specialized window that enables you to apply external methods to event streams. You can use it when complex procedural logic is required or when external methods or functions already exist. You can apply external methods by using C++, SAS DS2, SAS DATA step, or SAS® Micro Analytic Services. This example calls Python through SAS Micro Analytic Services; the code implements custom logic such as the following:

- o If sentiment is “Negative” and relevancy is close to 1, then recommend a sell.
- o If category is “Executive appointments” and sentiment is “Positive,” then recommend a buy.
- o If category is “Corporate reputation” and sentiment is “Positive,” then recommend a hold.

As events continuously flow into the system, a recommendation is assigned for each event. If the post is associated with negative sentiment, then the recommendation would be to sell the stock. Display 14 shows the output and recommendations that are generated by this window.



```
In [27]: procWindow
Out[27]:
```

		recommendation
term	_Index_	
$AAPL	1	HOLD
$AMZN	5	HOLD
$NFLX	10	BUY
$AMZN	10	BUY
	12	HOLD
	15	HOLD
$AAPL	15	HOLD
$QQQ	15	HOLD
$GOOGL	18	BUY
$GOOG	18	BUY
$AMZN	18	BUY
$AAPL	18	BUY

Display 14. Procedural Window Showing Final Recommendation for Each Event

OTHER APPLICATIONS

You can also use SAS Visual Text Analytics and SAS Event Stream Processing to address more mature business use cases, such as the following:

- Financial scenarios
  - o Quantitative investment and trading strategies: The trading and investment signals from real-time text analytics are applicable across all trading frequencies and provide an incremental source of quantitative factors.
  - o Algorithmic trading: You can enhance algorithmic strategies with automated circuit breakers, or you can develop new algorithms that take advantage of the ability to better predict trading volumes, price volatility, and directional movements.
  - o Market making: You can widen spreads or pull quotes when significant negative news is affecting a particular stock.
  - o Portfolio management: You can improve asset allocation decisions by benchmarking portfolio sentiment.
  - o Fundamental analysis: You can forecast stock, sector, and market outlooks.
- Non-financial scenarios
  - o Online email analysis of the mail exchange server to detect intellectual property (IP) leakage as emails are coming inbound and going outbound
  - o Fake-news detection and its possible impact on the stock market. Fake news can be identified in various ways: by examining the source, its popularity, and trustworthiness (Waldrop 2017).

CONCLUSION

This paper highlights how unstructured text analysis can be applied in-stream to provide a competitive advantage to financial technology institutions that use the analysis to drive algorithmic trading strategies. Although fintechs use more sophisticated algorithms, this approach demonstrates a simplified implementation that is very feasible within the framework of SAS Visual Text Analytics and SAS Event Stream Processing. This paper does not combine the results of structured data and unstructured text from tweets because access to real-time streaming sources was not available.

In-stream analytics occur as data streams from one device to another, or from multiple sensors to an aggregation point. Event stream processing is also supported at the "edge" so that you can analyze any data that are processed on the same device from which they are streaming.

REFERENCE

- Waldrop, M. Mitchell. 2017. "News Feature: The Genuine Problem of Fake News." *Proceedings of the National Academy of Sciences of the United States of America* 114 (48): 12631–12634. http://www.pnas.org/content/114/48/12631

ACKNOWLEDGMENTS

The authors thank Kevin Smith for helping implement the workflow using a Python interface to SAS Event Stream Processing.

They also thank Anne Baxter for editing the paper.

RECOMMENDED READING

- *SAS® Visual Text Analytics: Programming Guide*
- *SAS® Event Stream Processing: Programming Reference*
- Robert, Nicholas. "How to perform real time text analytics on Twitter streaming data in SAS ESP." Available https://blogs.sas.com/content/sgf/2016/10/05/how-to-perform-real-time-text-analytics-on-twitter-streaming-data-in-sas-esp/. Last modified October 5, 2016. Accessed on February 26, 2018.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Simran Bagga
SAS Institute Inc.
simran.bagga@sas.com

Saratendu Sethi
SAS Institute Inc.
saratendu.sethi@sas.com

Paper SAS2225-2018

Harvesting Unstructured Data to Reduce Anti-Money Laundering (AML) Compliance Risk

Austin Cook and Beth Herron, SAS Institute Inc.

ABSTRACT

As an anti-money laundering (AML) analyst, you face a never-ending job of staying one step ahead of nefarious actors (for example, terrorist organizations, drug cartels, and other money launderers). The financial services industry has called into question whether traditional methods of combating money laundering and terrorism financing are effective and sustainable. Heightened regulatory expectations, emphasis on 100% coverage, identification of emerging risks, and rising staffing costs are driving institutions to modernize their systems. One area gaining traction in the industry is to leverage the vast amounts of unstructured data to gain deeper insights. From suspicious activity reports (SARs) to case notes and wire messages, most financial institutions have yet to apply analytics to this data to uncover new patterns and trends that might not surface themselves in traditional structured data. This paper explores the potential use cases for text analytics in AML and provides examples of entity and fact extraction and document categorization of unstructured data using SAS® Visual Text Analytics.

INTRODUCTION

Financial Institutions dedicate substantial resources in support of government's efforts to curb money laundering and terrorism financing. Money laundering is the process of making funds that were gained through illegal channels appear legitimate, typically through a process of placement, layering, and integration. Terrorism financing is often more challenging to identify, as the funding can be raised through legitimate means, but later used to fund an act of terror or support a terrorist organization. Detecting these patterns can often feel like a game of "whack-a-mole;" by the time a new control is implemented to identify a known risk, the criminals have already changed their behavior to elude your efforts. The stakes are high, as the amount of money laundered per year is estimated to be 2 to 5% of global GDP. That's 2 trillion in USD according to the United Nations Office on Drugs and Crime (UNODC). In today's big-data environment, using modern technology to quickly identify financial crimes is critical.

A lot has changed globally since the early AML regimes of the 1970s. A growing regulatory landscape has led to higher penalties for program deficiencies. Banking has fundamentally changed with the creation of digital channels, faster payments, and new financial instruments. Data storage has become cheaper, opening the opportunity to process big data rapidly. Financial institutions have mostly adapted to these changes through enhancements to their rule-based detection programs and, as a result, have seen their headcount and costs soar. There's an appetite to overhaul the system to reduce false positive rates, increase the detection of money laundering, and automate many of the tedious tasks required in the investigations process. With the help of SAS® Visual Text Analytics, we can leverage artificial intelligence techniques to scale the human act of reading, organizing, and quantifying free-form text in meaningful ways, uncovering a rich source of underused risk data.

UNSTRUCTURED DATA SOURCES

While structured data such as transaction, account, and demographic information has been used in combating money laundering for years, financial institutions are just now beginning to see the value in harvesting unstructured data sources. These data sources are both vast and rich with valuable information that provides new data points, creates linkages, and identifies trends. Here is a list of the more notable sources of unstructured data that can be used for AML:

- **Wire Data -** Wire transfers between financial institutions contain much more valuable information than just the amount of money being sent. Along with origination, intermediary, and beneficiary data, wires often include free-form text including payment instructions and other messaging.

- **Transaction Review Memos** - The branch employees and client managers are the first line of defense when it comes to protecting the bank from money laundering. Typically, these individuals report valuable insight to the AML group through a transaction review memo. The details included in these memos are at the branch attendee’s discretion, but often they have supporting detail on why the transaction was deemed suspicious that might not be apparent in the transaction alone.
- **Case Data** - Anti-money laundering case data contains information enriched by the investigator during the life of the investigation. Cases generally contain several free-form text fields including notes, comments, and email correspondence as well as a narrative report explaining the final disposition. If suspicious activity is identified, a suspicious activity report (SAR) will be filed.
- **Suspicious Activity Report Data** - SARs are documents that financial institutions must file with their in-country government agency following the identification of potential unusual behavior related to money laundering or fraud. These documents are typically free-form text and generally contain several pieces of information about the person, company, and entity or entities of interest; the general findings from the investigator as to what the suspicious activity was; as well as any supporting evidence for the suspicious activity.
- **Negative News** - Beyond unstructured data your financial institution generates, there is a vast amount of publicly generated data from news and media organizations. Public news data can be used to identify supporting information about your customers including relationships to businesses or risky behaviors and criminal activity.
- **Email/Phone/Chat** - In addition to transactional data, risk factors might be identified in the non-transactional data stored by the bank in the form of email, phone, or chat conversations between the customer and employee.
- **Law Enforcement Requests** - Financial institutions have an obligation to identify subjects of law enforcement requests and file SARs where appropriate. Grand jury subpoenas, national security letters, and other requests are received in electronic format and contain text regarding persons of interest and requests for information.
- **Trade Documents** - The global trade system remains primarily a paper-based system. The trade documents (letters of credit, bills of lading, commercial invoices, other shipping documents) contain critical risk information in free-form text such as boycott language, dual use goods, inconsistent unit pricing, and other trade-based, money-laundering vulnerabilities.

USE CASES IN AML

Mining your unstructured data can be valuable in uncovering new insights to help combat money laundering in your financial institutions. Processing techniques such as theme detection, categorization, and entity or fact extraction are all ways to provide structure to free-form text. Once text is structured, there are several use cases to apply this data to ensure compliance:

- **Negative News Monitoring** - As an industry standard, financial institutions typically look for negative news related to high-risk customers and customers who have an open AML case. With the wide array of digital news made available daily, the identification of credible news can be challenging. Negative news not relevant to compliance can bias an investigator’s decision process, while missed news can leave an institution open to reputational risk. Coupled with bank policy and risk tolerance, an automated process to identify negative news and successfully link this information to customers provides both cost and time savings through automation.
- **Network Analytics** - Perhaps one of the best pieces of information for investigating AML is to understand relationships among your customers, as well as non-customers. Most institutions have structured data for known relationships among their customers, but often there are gaps with unknown relationships and those relationships with non-customers. Relationships and networks often surface through normal investigative procedures and are documented in case notes and SAR data. Storing this valuable information and displaying it for future use along with geographic tagging

provides deeper insights to the investigations process.

- **SAR Attribution Detection** - The detection of money laundering is an exercise in correctly identifying rare events in vast amounts of data. As the AML compliance industry starts to explore the application of artificial intelligence and machine learning to replace Boolean rules, the need for reliably labeled data (target variables) for training becomes even more important. Often, SARs are filed based on external information, but are attributed to the success of one or more rule-based scenarios. Text mining can help determine the correlation. This is critical to not only tune existing models, but also to allow banks to predict newly identified patterns in the future.
- **Trade Finance Document Categorization -** Deciphering trade documents is a tedious, manual process. We've been testing cognitive computing capabilities that are used for character recognition and natural language processing for document categorization. In a pilot with a tier 1 bank, our models read trade finance documents with ~99% accuracy and reduced the time to manually process the documents from several weeks to 26 seconds in an automated process.

EXAMPLE FRAMEWORK USING SAS® VISUAL TEXT ANALYTICS

This paper explores the process of processing unstructured data to support any of the use cases listed above. To demonstrate the potential applications, we will follow the framework below, primarily using SAS Visual Text Analytics as the enabling technology.

- **Data Acquisition –** Data is acquired for the example use case utilizing web scraping tools and is imported into SAS Visual Text Analytics.
- **Concept Extraction** – Predefined and customized concepts are generated to extract key facts from the unstructured data.
- **Text Parsing** – The individual records are parsed to enumerate the terms contained in the documents and apply filtering with start and stop lists.
- **Topic Generation** – Individual records are grouped into a collection of related themes containing similar subject matter automatically based on a bottom-up approach using the underlying terms.
- **Categorization** – Documents are classified into predetermined categories based on a top-down approach of the areas of interest using linguistic rules.
- **Post-Processing** – Output from SAS Visual Text Analytics is processed and prepared for use in modeling or investigative tools.

DATA ACQUISITION

While SAR information is not publicly available, we wanted to conduct our analysis on text data with similar content and format. The Internal Revenue Service (IRS) publishes summaries of significant money laundering cases each fiscal year, dating back to 2015. This data is rich with information, including people, organizations, risk typologies, locations, and other interesting data related to financial crimes. Below is an example of an IRS case from our data set:

"**Former Owners of Money Transmitter Business Sentenced for Conspiring to Structure Financial Transactions**
On October 25, 2016, in Scranton, Pennsylvania, German Ossa-Rocha was sentenced to 27 months in prison and two years of supervised release. On October 26, 2016, Mirela Desouza was sentenced to 18 months in prison and two years of supervised release. Ossa-Rocha and Desouza were the former owners of Tropical Express, a money transmitter service business located in Stroudsburg. Beginning in approximately January of 2008 and continuing through December 2011, Ossa-Rocha and Desouza structured financial transactions that represented the proceeds of drug trafficking in a manner intended to avoid reporting and recording requirements. The amount of funds involved in the structuring was approximately $340,000. The funds were transmitted by Ossa-Rocha and Desouza via wire transfers to the Dominican Republic." (IRS)

Web scraping tools were used to extract the various money laundering examples and write to a CSV file with four columns: observation number, year, title, and text narrative. The CSV file was then imported into SAS Visual Text Analytics for analysis.

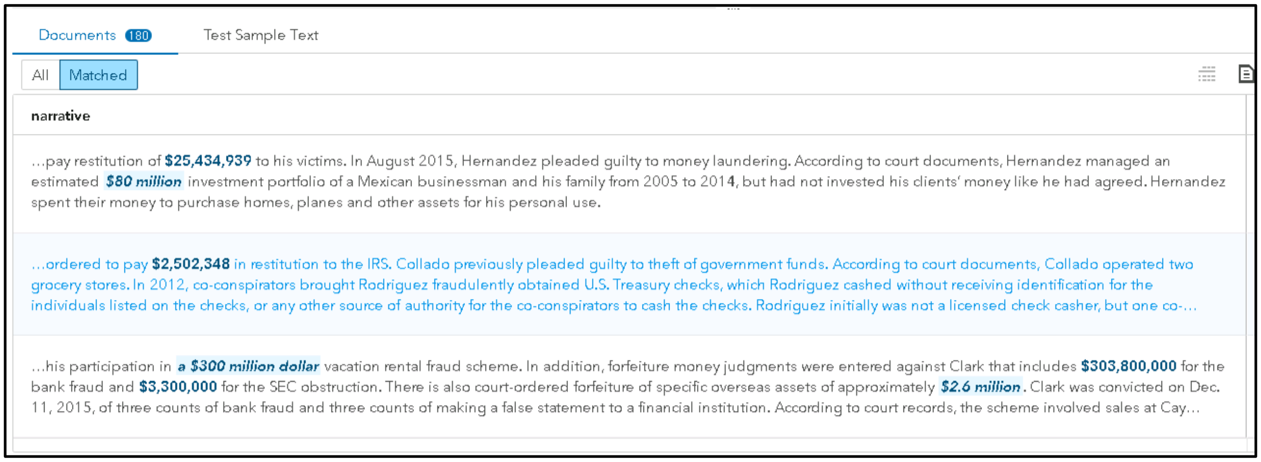
CONCEPT EXTRACTION

After initializing a project and loading the data, the first step in the process was focused on concept and fact extraction. With our data being rich in entities and facts, we wanted to extract these from the text for potential use in further analysis and research by investigators. In our model pipeline, this was done by dragging a Concept node and placing it on top of the Data node. SAS Visual Text Analytics comes with predefined concepts out of the box, as well as the ability to write your own custom concepts using LITI (language interpretation and text interpretation) syntax. For our analysis, we enabled the predefined concepts and wrote several custom concepts that are highlighted below.

The predefined concepts are common points of interest in which the rules come out of the box to immediately apply to your data, saving you time and helping you gain instant insights. Here are the predefined concepts of interest for our analysis:

- **nlpDate –** Identifies and extracts all dates and date ranges in your data in several formats (for example, May 2003, 05/15/2007, between 2007 and 2009, and so on).
- **nlpMeasure –** Identifies and extracts measures of time and quantities (for example, 30 years, 500 kilograms, and so on).
- **nlpMoney –** Identifies and extracts all references to currencies (for example, $272,000, more than $3 million, and so on).
- **nlpOrganizations –** Identifies and extracts all organization names (for example, U.S. Treasury, Department of Agriculture, and so on).
- **nlpPerson –** Identifies and extracts all names (for example, Joyce Allen, Robert L. Keys, and so on).
- **nlpPlace –** Identifies and extracts all places (for example, Asheville, North Carolina, Newport Beach, California, and so on).

Error! Reference source not found. below shows a set of matched concepts for the predefined concept nlpMoney.



narrative

...pay restitution of **$25,434,939** to his victims. In August 2015, Hernandez pleaded guilty to money laundering. According to court documents, Hernandez managed an estimated ***$80 million*** investment portfolio of a Mexican businessman and his family from 2005 to 2014, but had not invested his clients' money like he had agreed. Hernandez spent their money to purchase homes, planes and other assets for his personal use.

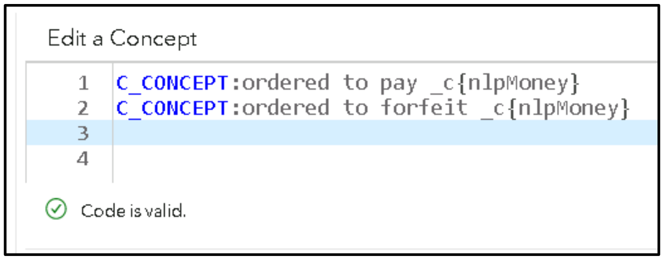
...ordered to pay **$2,502,348** in restitution to the IRS. Collado previously pleaded guilty to theft of government funds. According to court documents, Collado operated two grocery stores. In 2012, co-conspirators brought Rodriguez fraudulently obtained U.S. Treasury checks, which Rodriguez cashed without receiving identification for the individuals listed on the checks, or any other source of authority for the co-conspirators to cash the checks. Rodriguez initially was not a licensed check casher, but one co-...

...his participation in ***a $300 million dollar*** vacation rental fraud scheme. In addition, forfeiture money judgments were entered against Clark that includes **$303,800,000** for the bank fraud and **$3,300,000** for the SEC obstruction. There is also court-ordered forfeiture of specific overseas assets of approximately ***$2.6 million***. Clark was convicted on Dec. 11, 2015, of three counts of bank fraud and three counts of making a false statement to a financial institution. According to court records, the scheme involved sales at Cay...

Figure 1. Matched Concepts for Predefined Concept nlpMoney

While the predefined concepts are valuable in and of themselves, they are also useful for referencing in your custom concepts. An example of this can be seen with our custom concept Fine_Amount. The predefined concept nlpMoney will extract out all references to money, but suppose we want to exclusively extract out the fines associated with each record for further analysis. Instead of filtering through all references to money, we can define a custom concept to pull out only currencies associated with a fine.

Figure 2 below shows the LITI syntax to generate this rule:



```
C_CONCEPT:ordered to pay _c{nlpMoney}
C_CONCEPT:ordered to forfeit _c{nlpMoney}
```

Code is valid.

Figure 2. Custom Concept Fine_Amount LITI Syntax

The Fine_Amount custom concept uses the C_CONCEPT rule, which enables you to return matches that occur only in the context that we desire. In our case, we want to return the currency found by the nlpMoney predefined concept, but only in the context of a fine as in “ordered to pay” or “ordered to forfeit”.

A set of custom concepts was built on top of the predefined concepts to extract additional useful facts that could be helpful for indexing and searching, as well as additional analysis. Table 1 below summarizes the custom concepts that were developed, the type of concept used, and an example of the output.

Custom Concept	Concept Type	Example Output
Drug_Names	CLASSIFIER	Marijuana
Prison_Sentence	C_CONCEPT	60 months
Drug_Amount	CONCEPT_RULE	15 kilograms
Investment_Fraud_Amount	CONCEPT_RULE	$200 million
Investment_Fraud_Victims	CONCEPT_RULE	70 victims
Case_Charges	CLASSIFIER	Identity theft
Sentence_Location	CONCEPT_RULE	Providence, Rhode Island

Table 1. Custom Concept Definitions

TEXT PARSING

The next step in our analysis was to parse the text and create our term document matrix. In our model studio pipeline, this is done by dragging the Text Parsing node and placing it on top of the Concept node. SAS Visual Text Analytics allows you to customize how terms are parsed by configuring the minimum number of documents the term must be found in to be included for analysis, as well as using custom start, stop, and synonym lists. For the purposes of our example, we used the Text Parsing node to further explore some terms of interest for additional context and understanding. Figure 3 is an example of a term map used for exploration purposes.

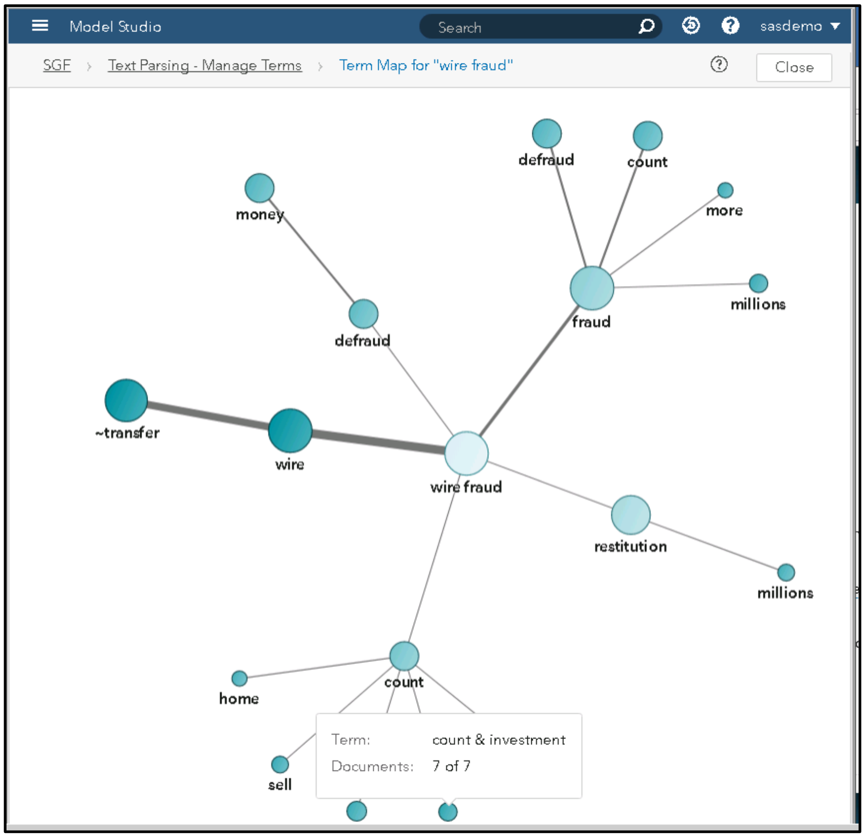
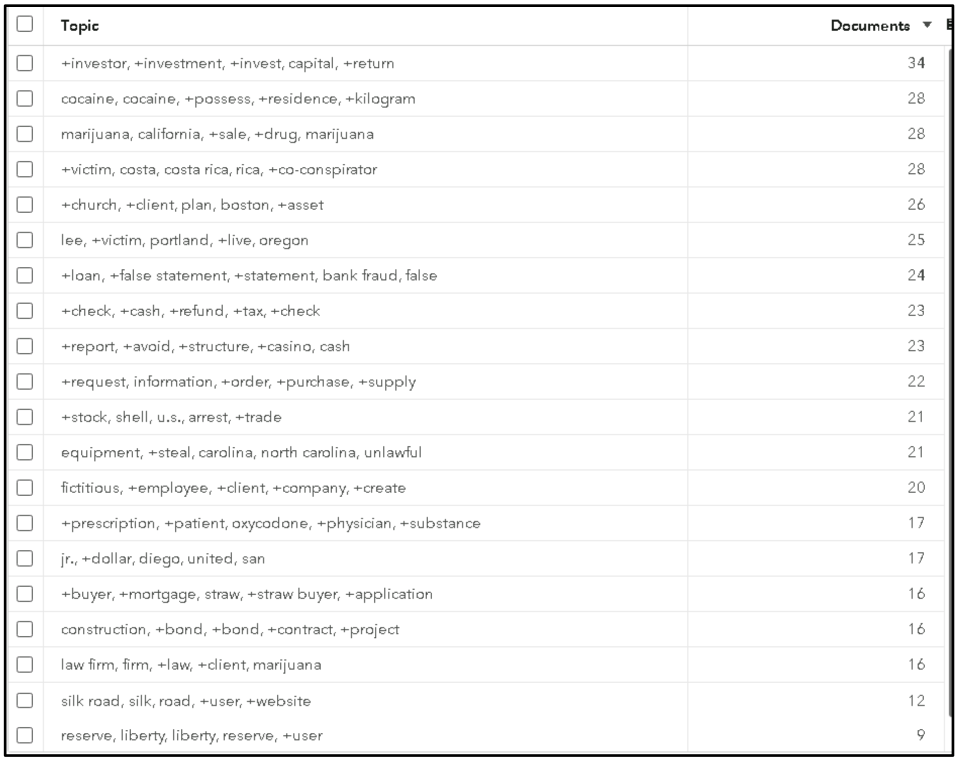


Figure 3. Term Map for "wire fraud"

TEXT TOPICS

Continuing with our analysis, we wanted to understand any relevant themes found in the data with the underlying terms that were parsed. For this, we dragged a Topic node and placed it on top of the Text Parsing node. SAS Visual Text Analytics allows you to automatically generate topics or choose the number of topics to generate, as well as set several other configurations including the term and document density. With a few iterations, we found the most informative results by setting the number of topics generated at 20, as well as term and document density of 2 and 1, respectively. Here is the output of the text topics.



☐	Topic	Documents ▼
☐	+investor, +investment, +invest, capital, +return	34
☐	cocaine, cocaine, +possess, +residence, +kilogram	28
☐	marijuana, california, +sale, +drug, marijuana	28
☐	+victim, costa, costa rica, rica, +co-conspirator	28
☐	+church, +client, plan, boston, +asset	26
☐	lee, +victim, portland, +live, oregon	25
☐	+loan, +false statement, +statement, bank fraud, false	24
☐	+check, +cash, +refund, +tax, +check	23
☐	+report, +avoid, +structure, +casino, cash	23
☐	+request, information, +order, +purchase, +supply	22
☐	+stock, shell, u.s., arrest, +trade	21
☐	equipment, +steal, carolina, north carolina, unlawful	21
☐	fictitious, +employee, +client, +company, +create	20
☐	+prescription, +patient, oxycodone, +physician, +substance	17
☐	jr., +dollar, diego, united, san	17
☐	+buyer, +mortgage, straw, +straw buyer, +application	16
☐	construction, +bond, +bond, +contract, +project	16
☐	law firm, firm, +law, +client, marijuana	16
☐	silk road, silk, road, +user, +website	12
☐	reserve, liberty, liberty, reserve, +user	9

Figure 4. Text Topics and Associated Document Count

Upon inspecting the topics, we were interested in two themes that were promoted to categories for ongoing analysis. The topics that were automatically generated provided a new lens on the data that we would like to track further and categorize new documents moving forward.

Topic Terms	Topic Theme	Percent of Documents
+buyer, +mortgage, straw, +straw buyer, +application	Real Estate Investment Fraud	9.4%
silk road, silk, road, +user, +website	Dark Web Drug Trade	7.0%

Table 2. Text Topics Promoted to Categories

TEXT CATEGORIES

Previously, we discussed text topics and the bottom-up approach of using the underlying terms to generate topics of interest. Our next step in our analysis was to take a top-down approach and define categories of interest using linguistic rules available in SAS Visual Text Analytics. In our model pipeline, this is done by dragging a Category node and placing it on top of the Topic node.

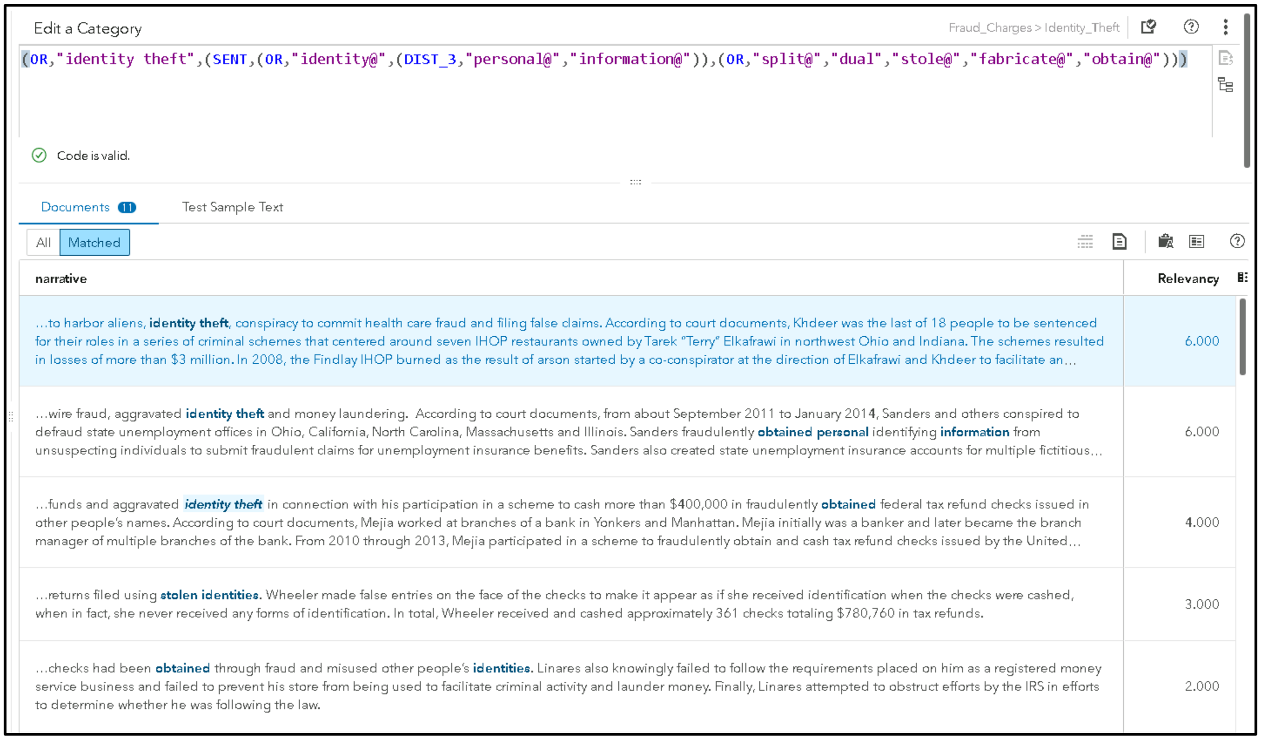
Categorizing your documents can be valuable for several reasons, such as creating tags for searching or for assigning similar documents for workflow purposes. Previously, we identified two categories of interest that we converted from the topics that were generated using the Topic node. In addition to these, we created a custom hierarchy of categorization that will help with future analysis. The table below shows the hierarchy of categories we were interested in.

Level 1	Level 2	Percentage of Matches
Drug Activity	Pharma Drugs	3%

	Illegal Drugs	15%
High Risk Customer Groups	Casino	3%
	Real Estate	23%
	Shell Company	3%
Financial Crime Charges	Bank Fraud	14%
	Bulk Cash Smuggling	4%
	Check Fraud	1%
	Identity Theft	6%
	Investment Fraud	8%
	Mail Fraud	16%
	Structuring	3%
	Tax Fraud	5%
	Wire Fraud	28%

Table 3. Custom Category Matches

Each category uses Boolean and proximity operators, arguments, and modifiers to effectively provide matches to only desired documents. Through the authors' domain expertise and the capabilities of SAS Visual Text Analytics, we were able to provide relevant matches on several categories of interest. An example of this concept is outlined below using the text category for the custom category "Identify Theft":



```
(OR,"identity theft",(SENT,(OR,"identity@",(DIST_3,"personal@","information@")),(OR,"split@","dual","stole@","fabricate@","obtain@")))
```

Code is valid.

narrative	Relevancy
...to harbor aliens, **identity theft**, conspiracy to commit health care fraud and filing false claims. According to court documents, Khdeer was the last of 18 people to be sentenced for their roles in a series of criminal schemes that centered around seven IHOP restaurants owned by Tarek "Terry" Elkafrawi in northwest Ohio and Indiana. The schemes resulted in losses of more than $3 million. In 2008, the Findlay IHOP burned as the result of arson started by a co-conspirator at the direction of Elkafrawi and Khdeer to facilitate an...	6.000
...wire fraud, aggravated **identity theft** and money laundering. According to court documents, from about September 2011 to January 2014, Sanders and others conspired to defraud state unemployment offices in Ohio, California, North Carolina, Massachusetts and Illinois. Sanders fraudulently **obtained personal** identifying **information** from unsuspecting individuals to submit fraudulent claims for unemployment insurance benefits. Sanders also created state unemployment insurance accounts for multiple fictitious...	6.000
...funds and aggravated ***identity theft*** in connection with his participation in a scheme to cash more than $400,000 in fraudulently **obtained** federal tax refund checks issued in other people's names. According to court documents, Mejia worked at branches of a bank in Yonkers and Manhattan. Mejia initially was a banker and later became the branch manager of multiple branches of the bank. From 2010 through 2013, Mejia participated in a scheme to fraudulently obtain and cash tax refund checks issued by the United...	4.000
...returns filed using **stolen identities**. Wheeler made false entries on the face of the checks to make it appear as if she received identification when the checks were cashed, when in fact, she never received any forms of identification. In total, Wheeler received and cashed approximately 361 checks totaling $780,760 in tax refunds.	3.000
...checks had been **obtained** through fraud and misused other people's **identities**. Linares also knowingly failed to follow the requirements placed on him as a registered money service business and failed to prevent his store from being used to facilitate criminal activity and launder money. Finally, Linares attempted to obstruct efforts by the IRS in efforts to determine whether he was following the law.	2.000

Figure 5. Text Category for "Identity Theft" with Matched Output

The "Identity Theft" rule can be broken up into two main components using the OR operator. The first component is simply looking for a direct match for the two sequential terms "identity theft", which provides several simple matches in the output found in the bottom of Figure 5. The second component uses the

SENT operator and will trigger a match if two sub-components exist in the same sentence somewhere within the document. The first sub-component is looking for some form of the word "identity" or a close combination of "personal" and "information". The second sub-component is looking for the action of theft including terms such as "split", "dual", "stole", "fabricate", or "obtain". The fourth and fifth matches in Figure 5 highlight the types of matches this will create in the form of "stolen identities" and "obtained identities" in the fourth and fifth match, respectively.

POST-PROCESSING

Once your project is set up in SAS Visual Text Analytics, you can produce score code and apply this to new data for ongoing tracking and monitoring. There are several types of post-processing that can happen depending on your use case and what the type of output you are working with. The most common types of post-processing can be found below:

- **Categorical Flags** – Typically, the presence or match for a category is used as a binary indicator for each document and can be used in filtering or searching, or as inputs to machine learning algorithms.
- **Network Analysis** – Extracted concepts such as locations, people, and organizations can be post-processed to show linkages and used as input to network diagrams for analysis.
- **Numerical Analysis** – Extracted concepts such as duration, fine amounts, or other numerical fields extracted from the documents can be post-processed to derive summarizations and averages of areas of interest.

CONCLUSION

There is a lot of excitement in the financial crime and compliance industry around the application of artificial intelligence and automation techniques. We see many opportunities available today to apply these methods to improve the effectiveness of detection programs and automate the manual tasks being performed by investigators. Text analytics is one area that has enormous potential, given that compliance departments have vast amounts of untapped, unstructured data sources. These sources contain rich information including who, where, what, when, and how that can be used as an input to many financial crimes use cases such as Negative News Monitoring, Trade Finance Monitoring, and SAR/STR Quality Assurance. With SAS Visual Text Analytics, banks can extract and derive meaning from text and organize it in a way that helps them perform these complex tasks that were previously accessible only through manual human review.

REFERENCES

UNODC (United Nations Office on Drugs and Crime). *n.d.* "Money-Laundering and Globalization." Accessed February 20, 2018. Available https://www.unodc.org/unodc/en/money-laundering/globalization.html.

IRS (Internal Revenue Service). 2017. "Examples of Money Laundering Investigations - Fiscal Year 2017." Accessed February 20, 2018. Available https://www.irs.gov/compliance/criminal-investigation/examples-of-money-laundering-investigations-for-fiscal-year-2017.

ACKNOWLEDGMENTS

The authors would like to thank David Stewart for his guidance and thought leadership regarding AML compliance. In addition, we would like to thank Adam Pilz for his guidance and thought leadership regarding text analytics.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Austin Cook
100 SAS Campus Drive
Cary, NC 27513

SAS Institute Inc.
Austin.Cook@sas.com
http://www.sas.com

Beth Herron
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Beth.Herron@sas.com
http://www.sas.com

SAS2257-2018

Invoiced: Using SAS® Contextual Analysis to Calculate Final Weighted Average Consumer Price

Alexandre Carvalho, SAS Institute Inc.

ABSTRACT

SAS® Contextual Analysis brings advantages to the analysis of the millions of Electronic Tax Invoices (Nota Fiscal Electrônica) issued by industries and improves the validation of taxes applied. Tax calculation is one of the analytical challenges for government finance secretaries in Brazil. This paper highlights two items of interest in the public sector: tax collection efficiency and the calculation of the final weighted average consumer price. The features in SAS® Contextual Analysis enable the implementation of a tax taxonomy that analyzes the contents of invoices, automatically categorizes the product, and calculates a reference value of the prices charged in the market. The first use case is an analysis of compliance between the official tax rate—as specified by the Mercosul Common Nomenclature (NCM)—and the description on the electronic invoice. (The NCM code was adopted in January 1995 by Argentina, Brazil, Paraguay, and Uruguay for product classification.) The second use case is the calculation of the final weighted average consumer price (PMPF). Generally, this calculation is done through sampling performed by public agencies. The benefits of a solution such as SAS Contextual Analysis are automatic categorization of all invoices and NCM code validation. The text analysis and the generated results contribute to tax collection efficiency and result in a more adequate reference value for use in the calculation of taxes on the circulation of goods and services.

INTRODUCTION

This paper focuses on the analytical challenges of government finance secretaries in Brazil, including the following:

- categorize the contents of the Electronic Tax Invoices
- improve the accuracy of the calculation of the final weighted average consumer price
- build an analytical base table that can be used as the basis for the calculation of the final weighted average consumer price

Business analysts and IT professionals are looking for solutions that are easy to use and easy to integrate into their existing systems, and that improve their analytics and their outcomes to challenges. SAS Contextual Analysis has benefits that combine machine learning and text mining with linguistic rules.

Some of these features can be directly monetized to help provide a fast return, such as the following:

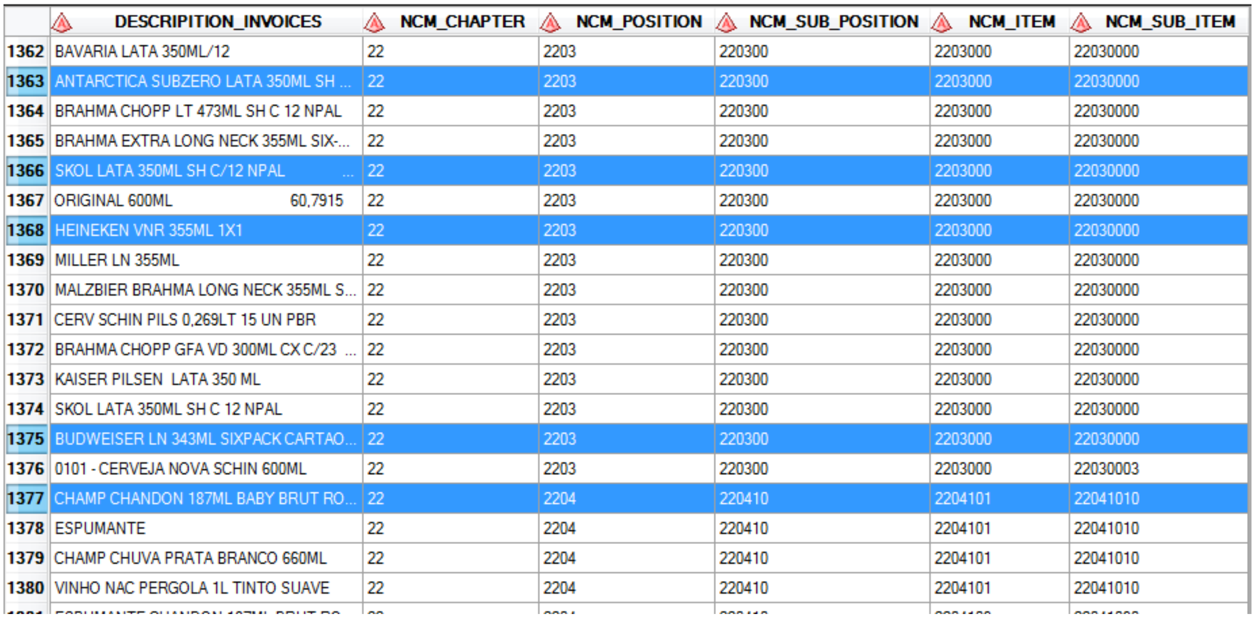
- filtering documents
- predefined concepts
- ability to create and improving rules to concepts and categories
- exploring for new topics
- categorizing unstructured textual data and collections of documents

These and other features are found in SAS Contextual Analysis through a single integrated system. You can update and customize rules as needed.

DATA SOURCES FOR THIS DEMO

The data source was provided by and its use authorized by Secretaria de Estado de Fazenda de Minas Gerais (SEFA MG), Brazil. In May 2017, the data source was utilized in Proof of Concept (POC) for categorizing invoice issues. The results were reduced classification time, improved accuracy in product identification, and help with identifying anomalies in invoices and taxes.

Display 1 shows a sample of the data source with 9,955 rows and 6 variables (including descriptive text about the invoices and the NCM code). The sample contains grouped information about Electronic Tax Invoices issued to taxpayers (that is, industries). The Electronic Tax Invoices issued are a selection of the invoices issued in May 2017, and the source does not contain confidential information about taxpayers.



	DESCRIPITION_INVOICES	NCM_CHAPTER	NCM_POSITION	NCM_SUB_POSITION	NCM_ITEM	NCM_SUB_ITEM
1362	BAVARIA LATA 350ML/12	22	2203	220300	2203000	22030000
1363	ANTARCTICA SUBZERO LATA 350ML SH ...	22	2203	220300	2203000	22030000
1364	BRAHMA CHOPP LT 473ML SH C 12 NPAL	22	2203	220300	2203000	22030000
1365	BRAHMA EXTRA LONG NECK 355ML SIX-...	22	2203	220300	2203000	22030000
1366	SKOL LATA 350ML SH C/12 NPAL ...	22	2203	220300	2203000	22030000
1367	ORIGINAL 600ML 60,7915	22	2203	220300	2203000	22030000
1368	HEINEKEN VNR 355ML 1X1	22	2203	220300	2203000	22030000
1369	MILLER LN 355ML	22	2203	220300	2203000	22030000
1370	MALZBIER BRAHMA LONG NECK 355ML S...	22	2203	220300	2203000	22030000
1371	CERV SCHIN PILS 0,269LT 15 UN PBR	22	2203	220300	2203000	22030000
1372	BRAHMA CHOPP GFA VD 300ML CX C/23 ...	22	2203	220300	2203000	22030000
1373	KAISER PILSEN LATA 350 ML	22	2203	220300	2203000	22030000
1374	SKOL LATA 350ML SH C 12 NPAL	22	2203	220300	2203000	22030000
1375	BUDWEISER LN 343ML SIXPACK CARTAO...	22	2203	220300	2203000	22030000
1376	0101 - CERVEJA NOVA SCHIN 600ML	22	2203	220300	2203000	22030003
1377	CHAMP CHANDON 187ML BABY BRUT RO...	22	2204	220410	2204101	22041010
1378	ESPUMANTE	22	2204	220410	2204101	22041010
1379	CHAMP CHUVA PRATA BRANCO 660ML	22	2204	220410	2204101	22041010
1380	VINHO NAC PERGOLA 1L TINTO SUAVE	22	2204	220410	2204101	22041010

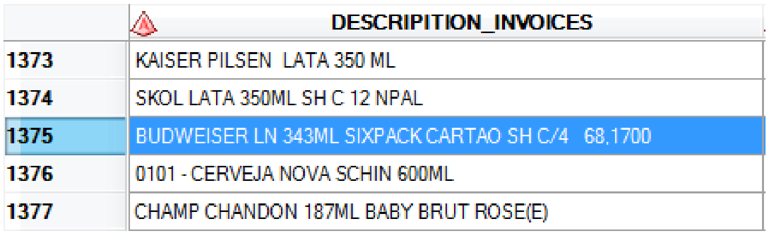
Display 1. Data Source from SEFA-MG, 2017

UNDERSTANDING THE ICMS TAX AND THE CONTENT OF THE INVOICE DESCRIPTIONS

ICMS is the tax levied on the circulation of products such as food, beverages, household appliances, communication services, transportation, and some imported products, and became law in 1997 (also known as the Lei Kandir law). In Brazil, ICMS is one of the largest sources of financial revenue. Because it is established by each state (for example, Minas Gerais, Rio de Janeiro, or São Paulo), it changes from one place to another. Tax collections can be routed to various functions (for example, health, education, payment of civil servants, and so on).

At each stage of the collection cycle, it is always necessary to issue an invoice or tax coupon, which is calculated by the taxpayer and collected by the State. There are two types of Electronic Tax Invoices: invoices issued at the industry level (electronic invoices issued by the beer, refrigerator, or fuel industries) and invoices issued at the consumer level (electronic invoices issued by restaurants to final consumers).

In Display 2, line 1375 (BUDWEISER LN 343ML SIXPACK CARTAO) provides us with the following information: Product (Budweiser), Type (LN means Long Neck), Volume (343ML), and Quantity (SIXPACK CARTAO SH C/4).



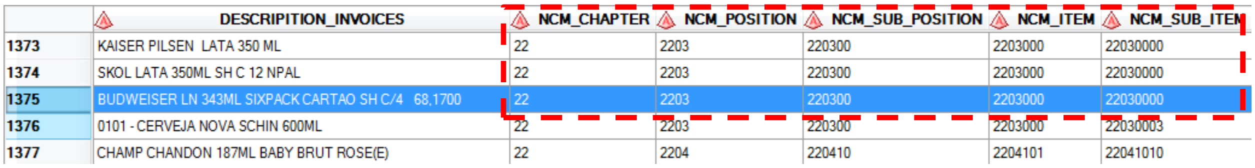
	DESCRIPITION_INVOICES
1373	KAISER PILSEN LATA 350 ML
1374	SKOL LATA 350ML SH C 12 NPAL
1375	BUDWEISER LN 343ML SIXPACK CARTAO SH C/4 68,1700
1376	0101 - CERVEJA NOVA SCHIN 600ML
1377	CHAMP CHANDON 187ML BABY BRUT ROSE(E)

Display 2. Data Source Content

WHAT IS THE MERCOSUL COMMON NOMENCLATURE (NCM CODE) FOR PRODUCT CLASSIFICATION?

The classification system for invoices follows the Mercosul Common Nomenclature (Nomenclatura Comum do Mercosul, or NCM) and was adopted in January 1995 by Argentina, Brazil, Paraguay, and Uruguay for product classification. Any merchandise, imported or purchased in Brazil, must have an NCM code in its legal documentation (invoices, legal books, and so on), whose objective is to classify the items according to the Mercosul regulation.

Display 3 shows examples of the content of Electronic Tax Invoices according to the NCM code by chapter, position, sub-position, item, and sub-item.



	DESCRIPITION_INVOICES	NCM_CHAPTER	NCM_POSITION	NCM_SUB_POSITION	NCM_ITEM	NCM_SUB_ITEM
1373	KAISER PILSEN LATA 350 ML	22	2203	220300	2203000	22030000
1374	SKOL LATA 350ML SH C 12 NPAL	22	2203	220300	2203000	22030000
1375	BUDWEISER LN 343ML SIXPACK CARTAO SH C/4 68,1700	22	2203	220300	2203000	22030000
1376	0101 - CERVEJA NOVA SCHIN 600ML	22	2203	220300	2203000	22030003
1377	CHAMP CHANDON 187ML BABY BRUT ROSE(E)	22	2204	220410	2204101	22041010

Display 3. Mercosul Common Nomenclature Content

IMPROVING CATEGORIZATION EFFICIENCY WITH SAS CONTEXTUAL ANALYSIS

The use of unstructured data is growing exponentially in government agencies. In January 2018, according to the Brazilian Federal Revenue Agency (Receita Federal Brasileira), approximately 18 billion Electronic Tax Invoices were identified, and the number of issuers was approximately 1.4 million.

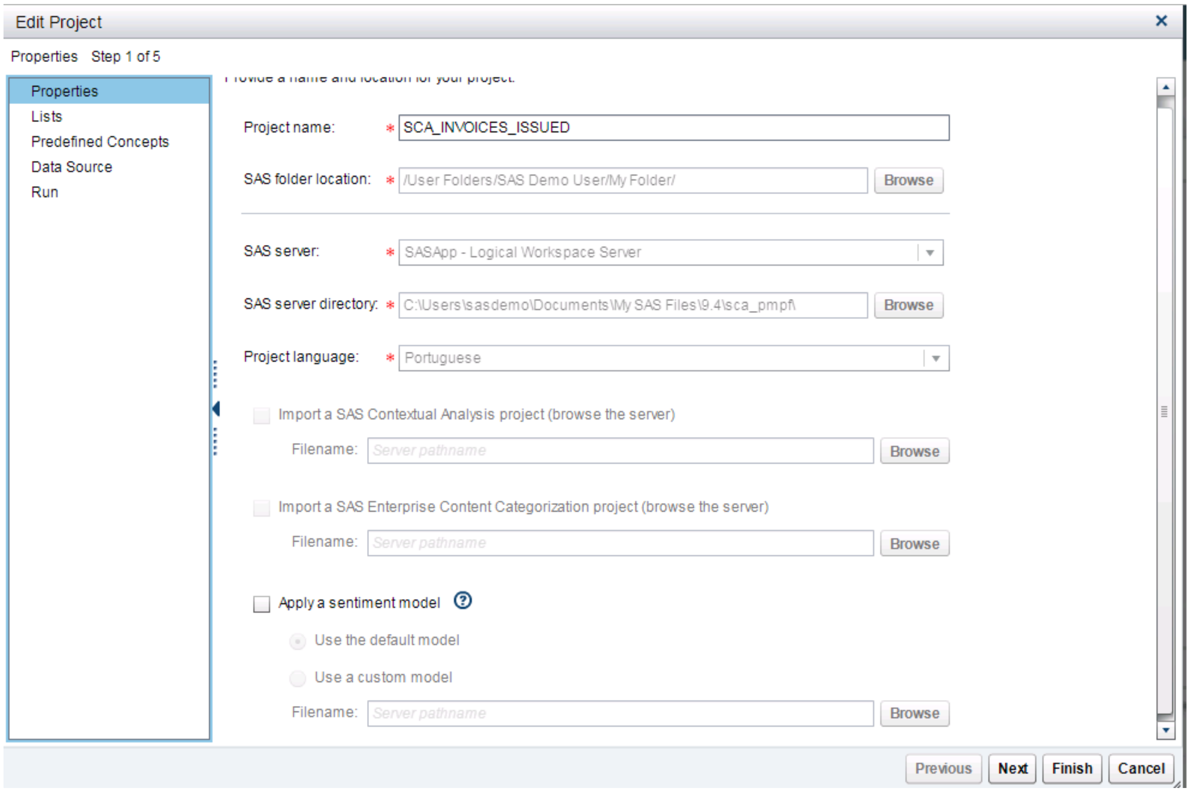
THE BENEFITS OF USING SAS CONTEXTUAL ANALYSIS

Business analysts are looking for solutions that are fast, easy to use and integrate into existing systems, and that improve their analytics and challenges. For the classification of electronic invoices, the analyst has more control with a hybrid approach. Analysts can add concepts (for example, 1LT, 500GR means quantity) and synonyms (skol, Budweiser, heinecken, brhama means beer) that specifically identify the product and its value for the tax aliquot calculation (for example, beer and 1LT the tax aliquot is 4%).

SAS Contextual Analysis combines machine learning and text mining capabilities with the ability to impose linguistic rules. SAS Contextual Analysis also enables you to filter, explore, and categorize unstructured textual data and collections of documents.Technology syntactically identifies common themes, category rules, and document sentiment, based on data. At any time, you can review and modify the results to meet your specific needs.

HOW TO BUILD A PROJECT IN SAS CONTEXTUAL ANALYSIS

Display 4 shows Step 1 of 5 for building a project in SAS Contextual Analysis.The analyst defines the name and location for your project, and chooses a project language. This paper doesn't apply a sentiment model, but is possible to use either the default model or a custom model.



Properties Step 1 of 5

Provide a name and location for your project.

Project name: * SCA_INVOICES_ISSUED

SAS folder location: * /User Folders/SAS Demo User/My Folder/

SAS server: * SASApp - Logical Workspace Server

SAS server directory: * C:\Users\sasdemo\Documents\My SAS Files\9.4\sca_pmpf\

Project language: * Portuguese

□ Import a SAS Contextual Analysis project (browse the server)

Filename: Server pathname

□ Import a SAS Enterprise Content Categorization project (browse the server)

Filename: Server pathname

□ Apply a sentiment model

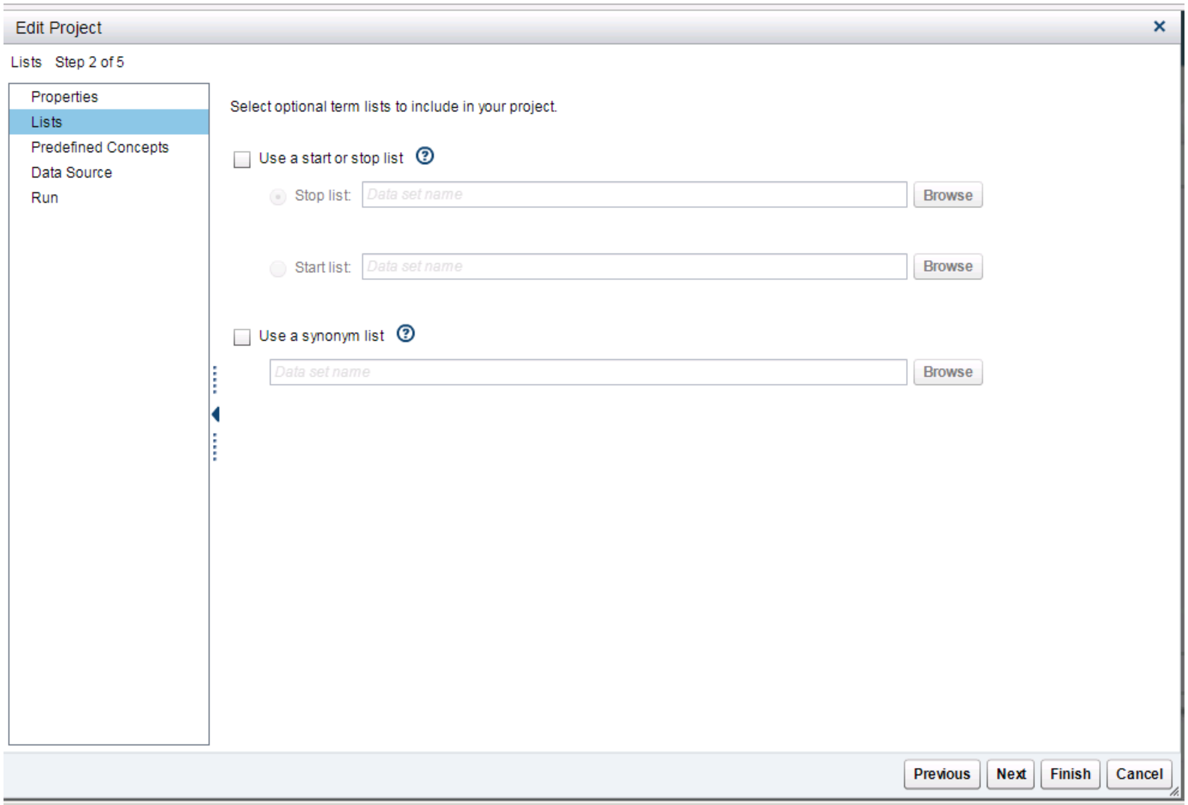
● Use the default model

○ Use a custom model

Filename: Server pathname

Display 4. Create a New Project: Define name, location and language for your project

Display 5 shows Step 2 of 5 for building a project in SAS Contextual Analysis. When analyzing text, it is common to disregard some terms already known to analysts that would not add value to the analysis or select a list of terms for research. For example, we can use the stop list (for name Brazil, SEFA-MG) or start list (skol, brahma, or budweiser). Another important feature is to use a list of synonyms whose terms would have the same meaning across the business (LT, GR, and KG all indicate quantity).



Select optional term lists to include in your project.

□ Use a start or stop list

● Stop list: Data set name

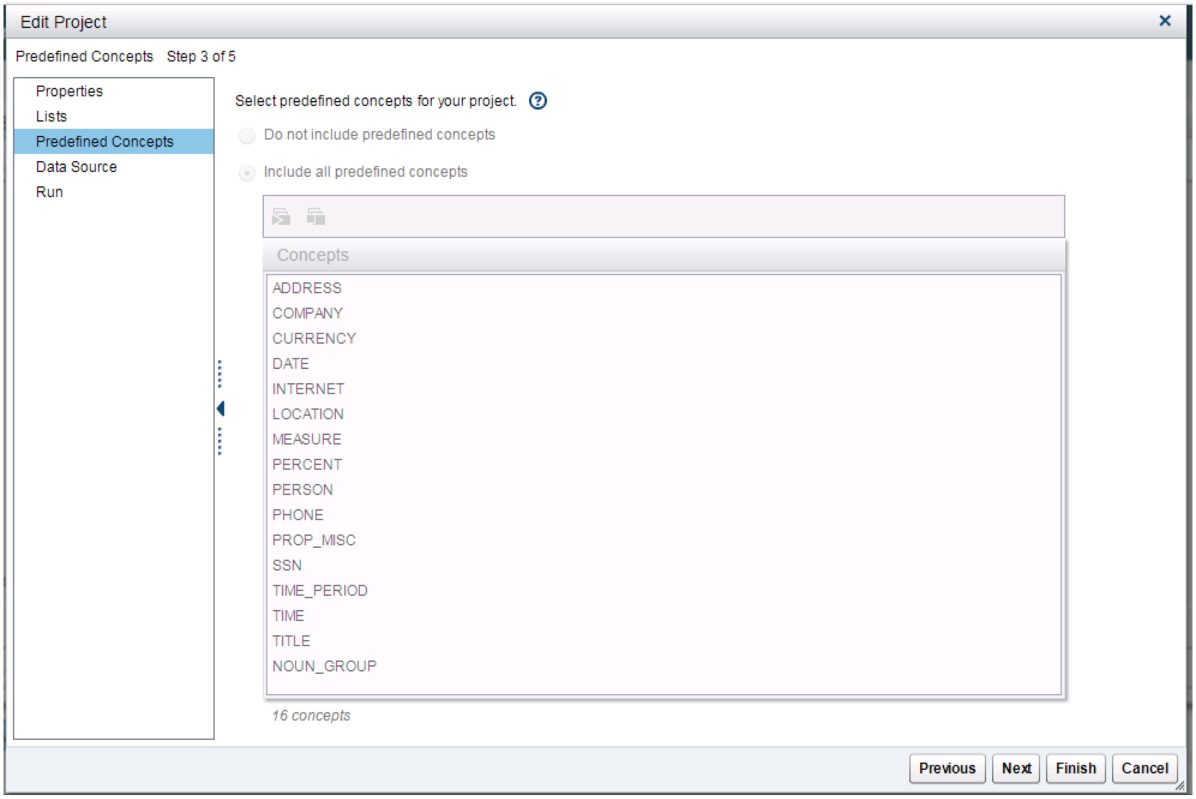
○ Start list: Data set name

□ Use a synonym list

Data set name

Display 5. Create a New Project: Define start list, stop list or synonyms list

Display 6 shows predefined concepts for your analysis and how SAS Contextual Analysis automatically identifies concepts such as location, currency, company, address, and so on.



Predefined Concepts Step 3 of 5

Select predefined concepts for your project.

○ Do not include predefined concepts

● Include all predefined concepts

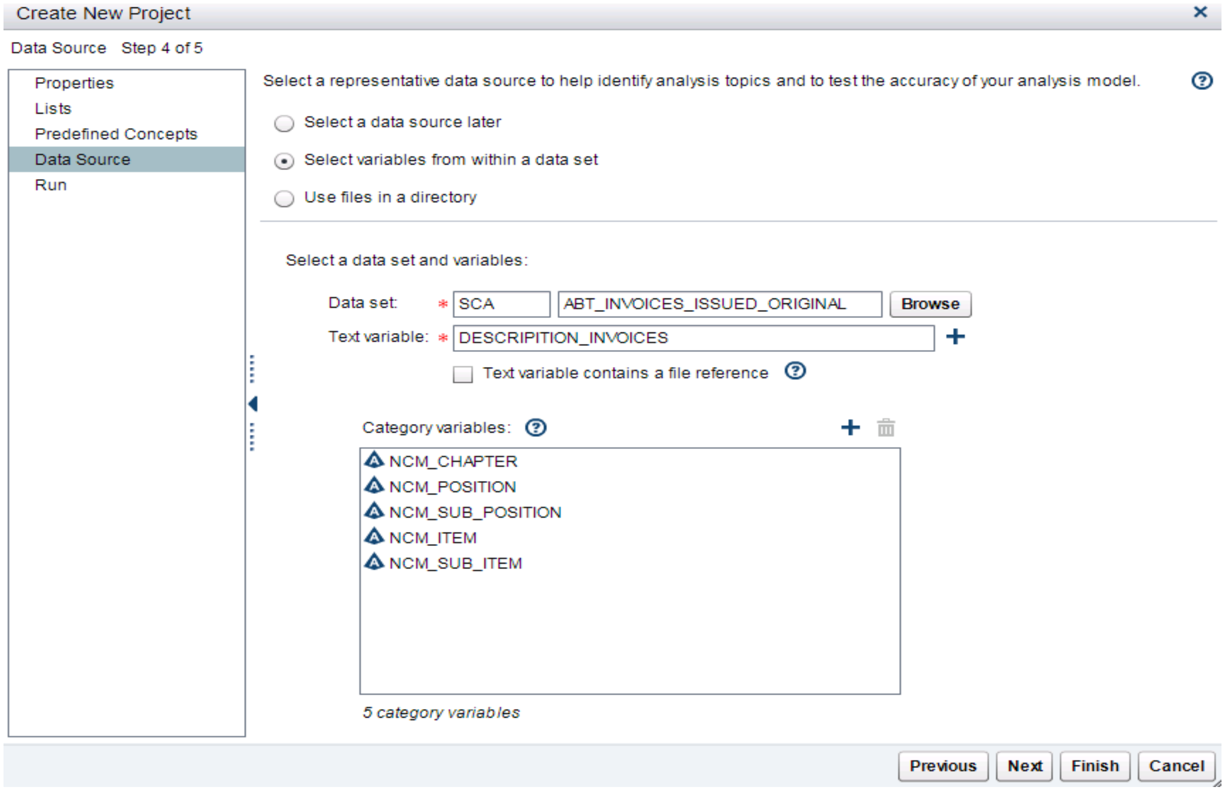
Concepts

ADDRESS
COMPANY
CURRENCY
DATE
INTERNET
LOCATION
MEASURE
PERCENT
PERSON
PHONE
PROP_MISC
SSN
TIME_PERIOD
TIME
TITLE
NOUN_GROUP

16 concepts

Display 6. Create a New Project: Predefined Concepts

Display 7 shows Step 4 of 5, which is when you select a SAS data set (ABT_INVOICES_ISSUED_ORIGINAL). The variable DESCRIPTION_INVOICES contains the invoice description, and text mining is used. On the other hand, NCM code information is used for categorization.


Create New Project ×

Data Source Step 4 of 5

Select a representative data source to help identify analysis topics and to test the accuracy of your analysis model.

○ Select a data source later

● Select variables from within a data set

○ Use files in a directory

Select a data set and variables:

Data set: * SCA ABT_INVOICES_ISSUED_ORIGINAL Browse

Text variable: * DESCRIPITION_INVOICES

□ Text variable contains a file reference

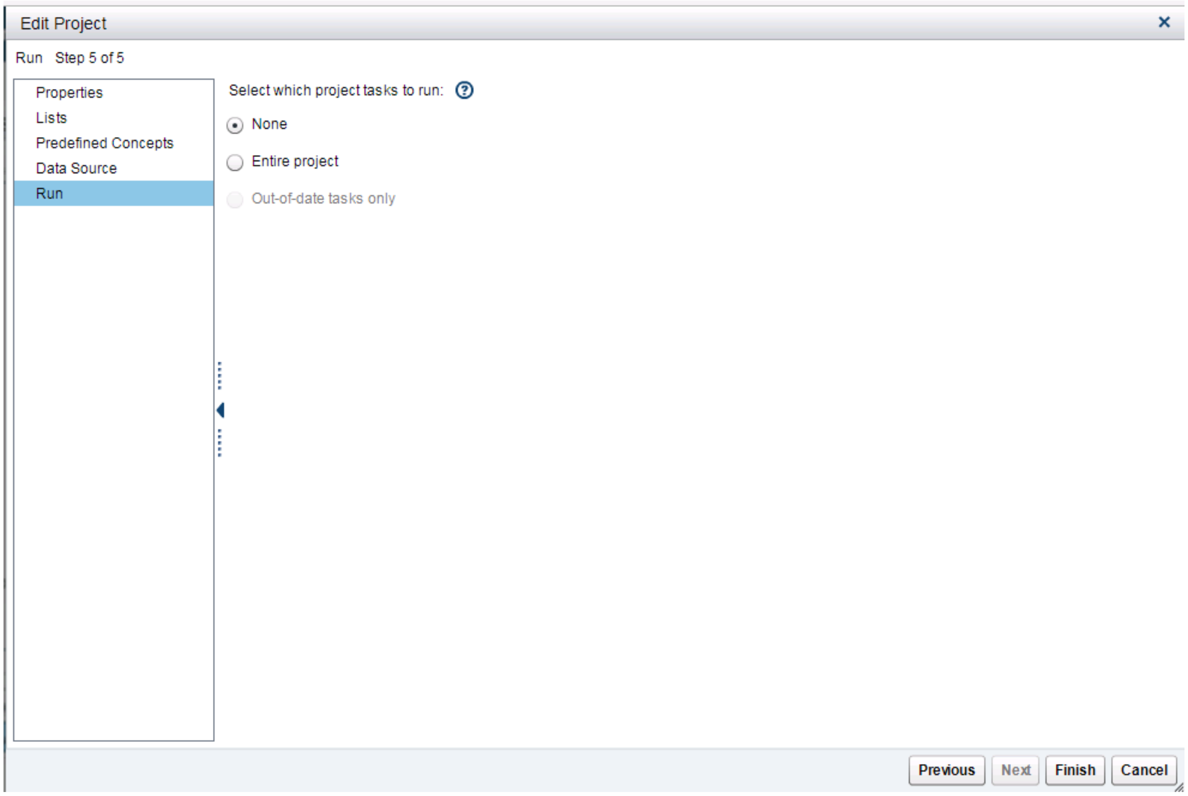
Category variables:

- NCM_CHAPTER
- NCM_POSITION
- NCM_SUB_POSITION
- NCM_ITEM
- NCM_SUB_ITEM

5 category variables

Display 7. Create a New Project: Select Data Set and Variables

And finally, you are ready to run the project (Display 8).



Edit Project

Run Step 5 of 5

- Properties
- Lists
- Predefined Concepts
- Data Source
- Run

Select which project tasks to run:

● None

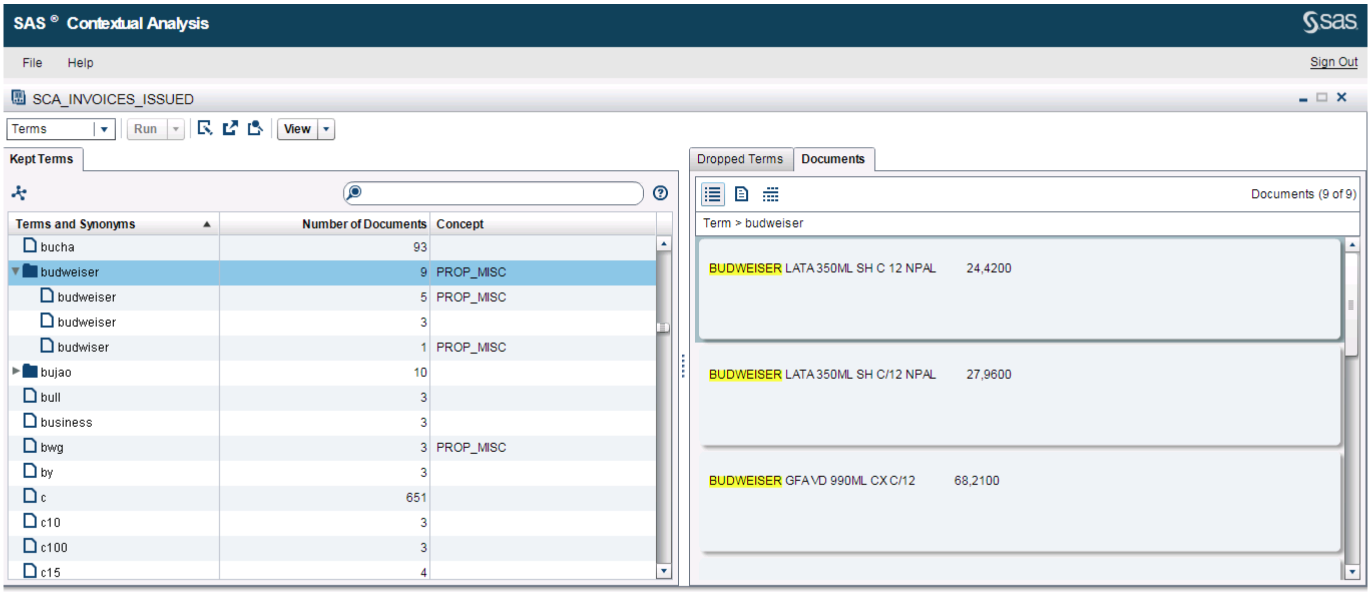
○ Entire project

○ Out-of-date tasks only

Display 8. Create a New Project: Run the entire project

IDENTIFY TERMS: NAME, TYPE, AND PRODUCT QUANTITY

Display 9 focuses on the term “budweiser”. In this case, you can see the stemming for the term “budweiser”, including the three forms it takes and the few rare misspellings that have occurred in the documents (for example, “budwiser”). In this example, "budweiser" is the description of a type of beer (product name).



SCA_INVOICES_ISSUED

Terms and Synonyms	Number of Documents	Concept
bucha	93	
budweiser	9	PROP_MISC
budweiser	5	PROP_MISC
budweiser	3	
budwiser	1	PROP_MISC
bujao	10	
bull	3	
business	3	
bwg	3	PROP_MISC
by	3	
c	651	
c10	3	
c100	3	
c15	4	

Documents (9 of 9)

Term > budweiser

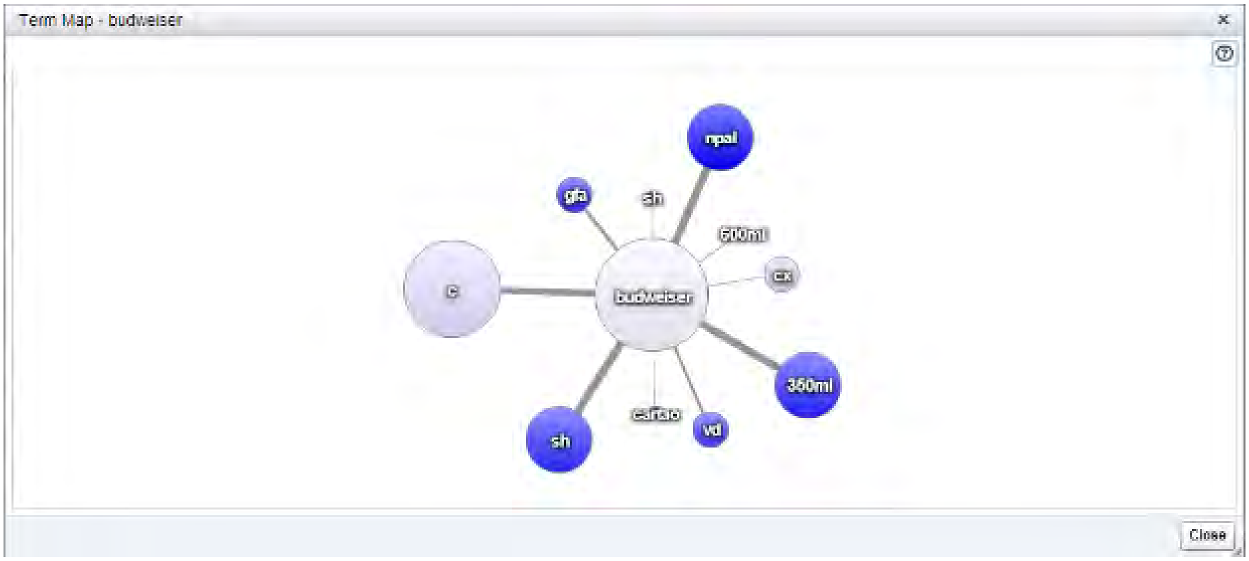
BUDWEISER LATA 350ML SH C 12 NPAL 24,4200

BUDWEISER LATA 350ML SH C/12 NPAL 27,9600

BUDWEISER GFA VD 990ML CX C/12 68,2100

Display 9. Create a New Project: Identifying Terms

In the term map shown in Display 10, you can see that there is additional information about the product type (for example, “gf” and “cx” mean “bottle”) and volume (350ml or 600ml). The term map can help you refine your terms list and create rules for classification.

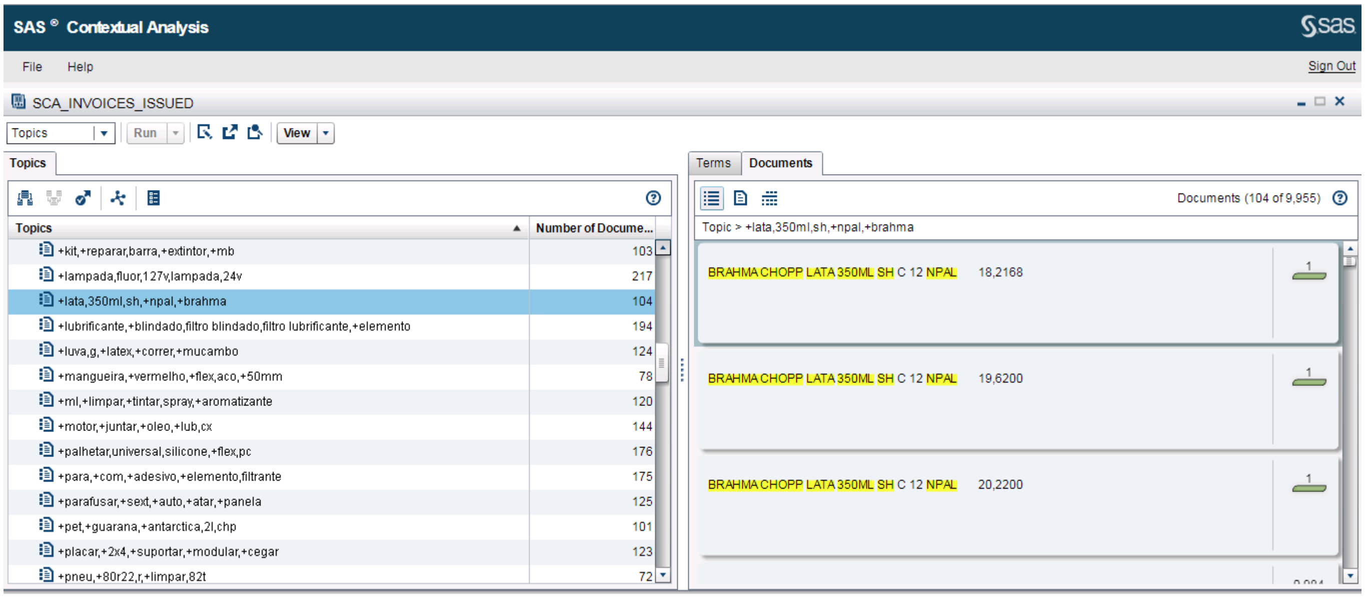


Display 10. Create New Project: Term Map

DISCOVERING TOPICS FOR THE ELECTRONIC TAX INVOICES

In particular, the Topics functionality in SAS Contextual Analysis can help you to automatically identify the contents of of your documents, which are in this case Electronic Tax Invoices.

Display 11 shows the documents for the topic **+lata+350ml,sh,+npal+brhama** . On the right side of the window, you can see a set of tax invoices that identify as a type of beer.



Topics	Number of Docume...
+kit,+reparar,barra,+extintor,+mb	103
+lampada,fluor,127v,lampada,24v	217
+lata,350ml,sh,+npal,+brahma	104
+lubrificante,+blindado,filtro blindado,filtro lubrificante,+elemento	194
+luva,g,+latex,+correr,+mucambo	124
+mangueira,+vermelho,+flex,aco,+50mm	78
+ml,+limpar,+tintar,spray,+aromatizante	120
+motor,+juntar,+oleo,+lub,cx	144
+palhetar,universal,silicone,+flex,pc	176
+para,+com,+adesivo,+elemento,filtrante	175
+parafusar,+sext,+auto,+atar,+panela	125
+pet,+guarana,+antarctica,2l,chp	101
+placar,+2x4,+suportar,+modular,+cegar	123
+pneu,+80r22,r,+limpar,82t	72

Topic > +lata,350ml,sh,+npal,+brahma

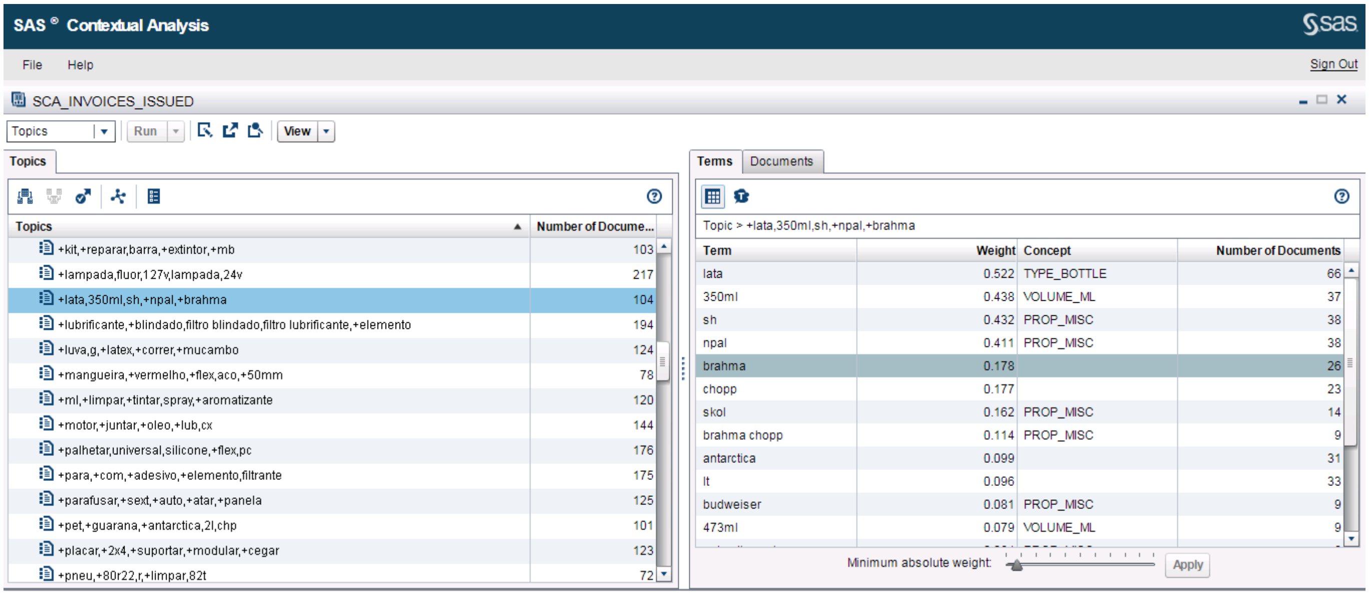
BRAHMA CHOPP LATA 350ML SH C 12 NPAL 18,2168 1

BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200 1

BRAHMA CHOPP LATA 350ML SH C 12 NPAL 20,2200 1

Display 11. Identify Emerging Issues

In Display 12 and Display 13, you see the **Terms** tab, on which you can choose from two different views of the terms that constitute the topics. You can also choose different views of the documents that are associated with the topics.

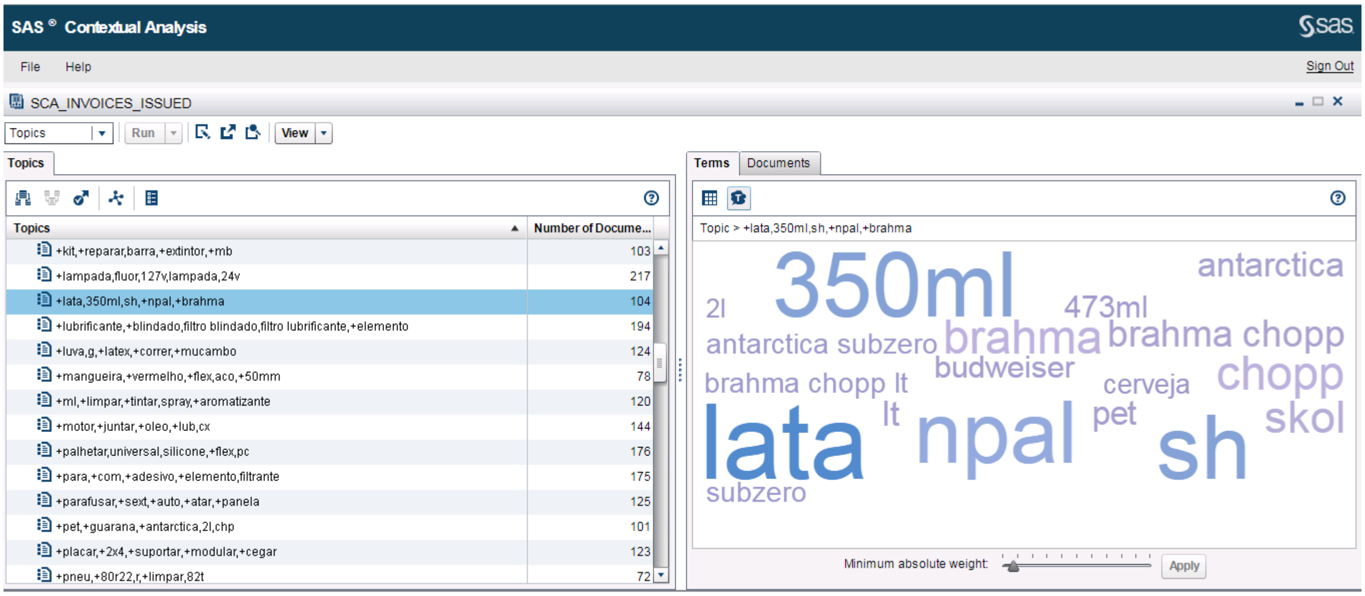


Topics	Number of Docume...
+kit,+reparar,barra,+extintor,+mb	103
+lampada,fluor,127v,lampada,24v	217
+lata,350ml,sh,+npal,+brahma	104
+lubrificante,+blindado,filtro blindado,filtro lubrificante,+elemento	194
+luva,g,+latex,+correr,+mucambo	124
+mangueira,+vermelho,+flex,aco,+50mm	78
+ml,+limpar,+tintar,spray,+aromatizante	120
+motor,+juntar,+oleo,+lub,cx	144
+palhetar,universal,silicone,+flex,pc	176
+para,+com,+adesivo,+elemento,filtrante	175
+parafusar,+sext,+auto,+atar,+panela	125
+pet,+guarana,+antarctica,2l,chp	101
+placar,+2x4,+suportar,+modular,+cegar	123
+pneu,+80r22,r,+limpar,82t	72

Topic > +lata,350ml,sh,+npal,+brahma

Term	Weight	Concept	Number of Documents
lata	0.522	TYPE_BOTTLE	66
350ml	0.438	VOLUME_ML	37
sh	0.432	PROP_MISC	38
npal	0.411	PROP_MISC	38
brahma	0.178		26
chopp	0.177		23
skol	0.162	PROP_MISC	14
brahma chopp	0.114	PROP_MISC	9
antarctica	0.099		31
lt	0.096		33
budweiser	0.081	PROP_MISC	9
473ml	0.079	VOLUME_ML	9

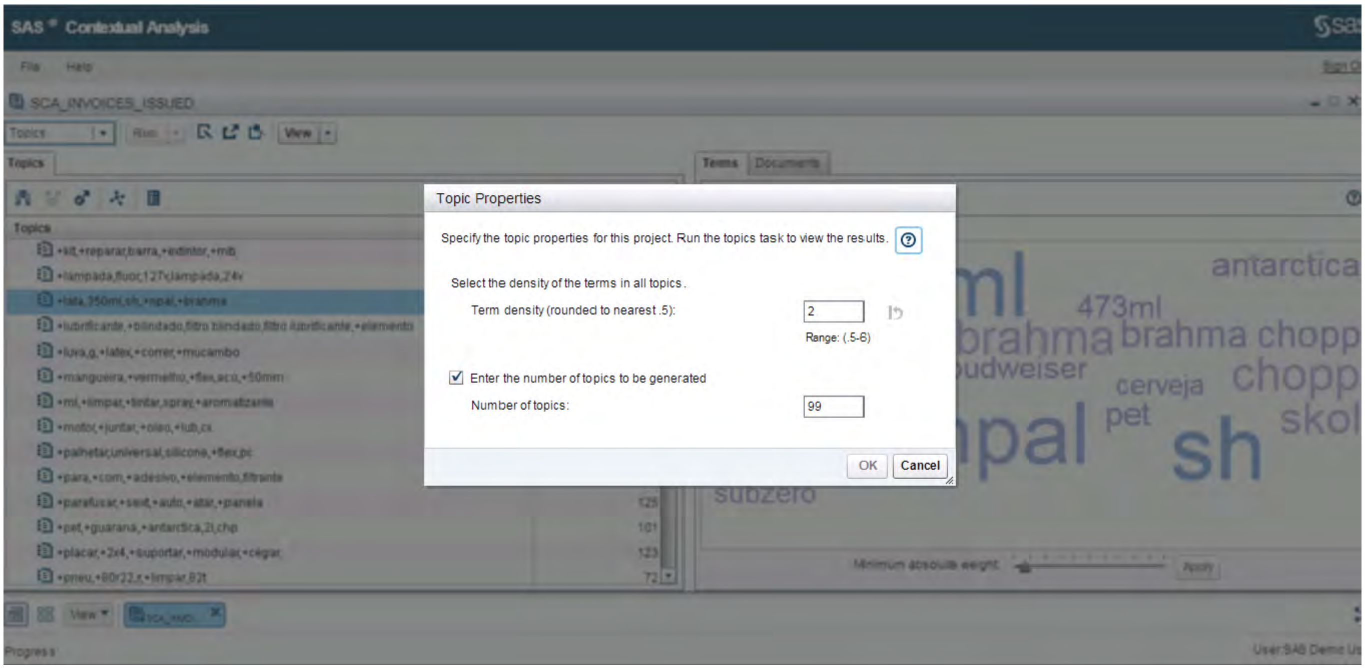
Display 12. The Terms Tab: View Tabular Form



Topics	Number of Docume...
+kit,+reparar,barra,+extintor,+mb	103
+lampada,fluor,127v,lampada,24v	217
+lata,350ml,sh,+npal,+brahma	104
+lubrificante,+blindado,filtro blindado,filtro lubrificante,+elemento	194
+luva,g,+latex,+correr,+mucambo	124
+mangueira,+vermelho,+flex,aco,+50mm	78
+ml,+limpar,+tintar,spray,+aromatizante	120
+motor,+juntar,+oleo,+lub,cx	144
+palhetar,universal,silicone,+flex,pc	176
+para,+com,+adesivo,+elemento,filtrante	175
+parafusar,+sext,+auto,+atar,+panela	125
+pet,+guarana,+antarctica,2l,chp	101
+placar,+2x4,+suportar,+modular,+cegar	123
+pneu,+80r22,r,+limpar,82t	72

Display 13. The Terms Tab: View Graphic Form

In some situations, the analyst needs to define a specific number of topics because of the structure of their challenges. In Display 14, we change the number of the topics to 99.



Topic Properties

Specify the topic properties for this project. Run the topics task to view the results.

Select the density of the terms in all topics.

Term density (rounded to nearest .5): 2

Range: (.5-6)

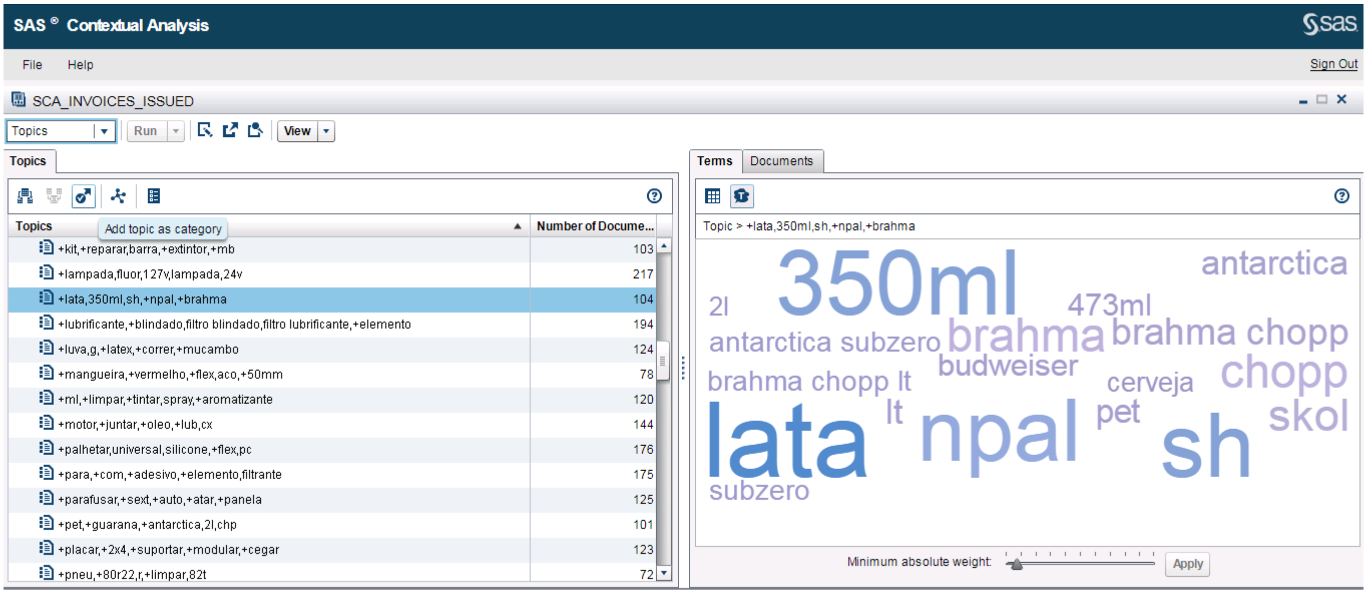
☑ Enter the number of topics to be generated

Number of topics: 99

Display 14. Topic Properties

HOW TO TRANSFORM TOPICS INTO CATEGORIES

After the topics are validated, you can create categories. Let's continue with the topic that identifies drinks, and promote some topics to be categories. First, you choose a topic and click the Add Topic icon, as shown in Display 15.



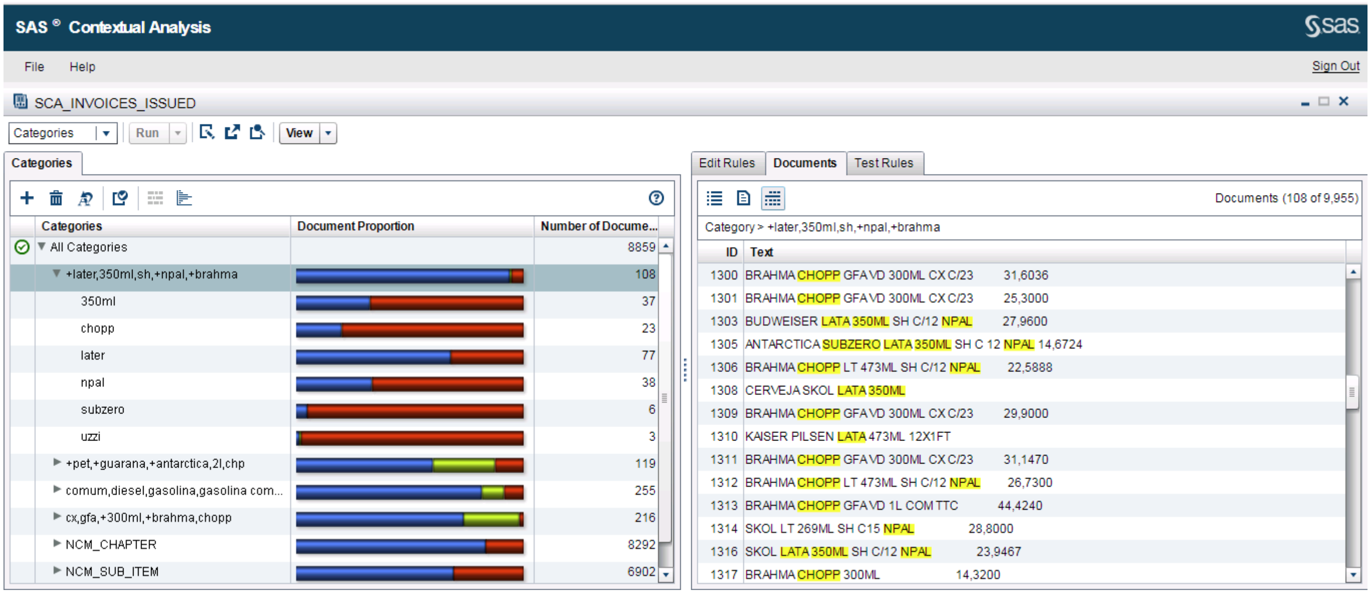
Topics	Number of Docume...
+kit,+reparar,barra,+extintor,+mb	103
+lampada,fluor,127v,lampada,24v	217
+lata,350ml,sh,+npal,+brahma	104
+lubrificante,+blindado,filtro blindado,filtro lubrificante,+elemento	194
+luva,g,+latex,+correr,+mucambo	124
+mangueira,+vermelho,+flex,aco,+50mm	78
+ml,+limpar,+tintar,spray,+aromatizante	120
+motor,+juntar,+oleo,+lub,cx	144
+palhetar,universal,silicone,+flex,pc	176
+para,+com,+adesivo,+elemento,filtrante	175
+parafusar,+sext,+auto,+atar,+panela	125
+pet,+guarana,+antarctica,2l,chp	101
+placar,+2x4,+suportar,+modular,+cegar	123
+pneu,+80r22,r,+limpar,82t	72

Topic > +lata,350ml,sh,+npal,+brahma

Display 15. Promote Topics to Categories

SAS Contextual Analysis suggests possible rules for classifying newly issued invoices. In this example, we transform the topic, which is the type of drinks, into a category that is defined as BEERS (see Display 16). On the **Documents** tab, you can see that out of 9,955 documents, 108 were categorized belonging to the BEERS category

This analysis evaluates how well the displayed linguistic definitions of the categories approximate the underlying machine learning definitions. This is important because you will use the linguistic definitions to score other documents.



Categories	Document Proportion	Number of Docume...
All Categories		8859
+later,350ml,sh,+npal,+brahma		108
350ml		37
chopp		23
later		77
npal		38
subzero		6
uzzi		3
+pet,+guarana,+antarctica,2l,chp		119
comum,diesel,gasolina,gasolina com...		255
cx,gfa,+300ml,+brahma,chopp		216
NCM_CHAPTER		8292
NCM_SUB_ITEM		6902

Documents (108 of 9,955)

Category > +later,350ml,sh,+npal,+brahma

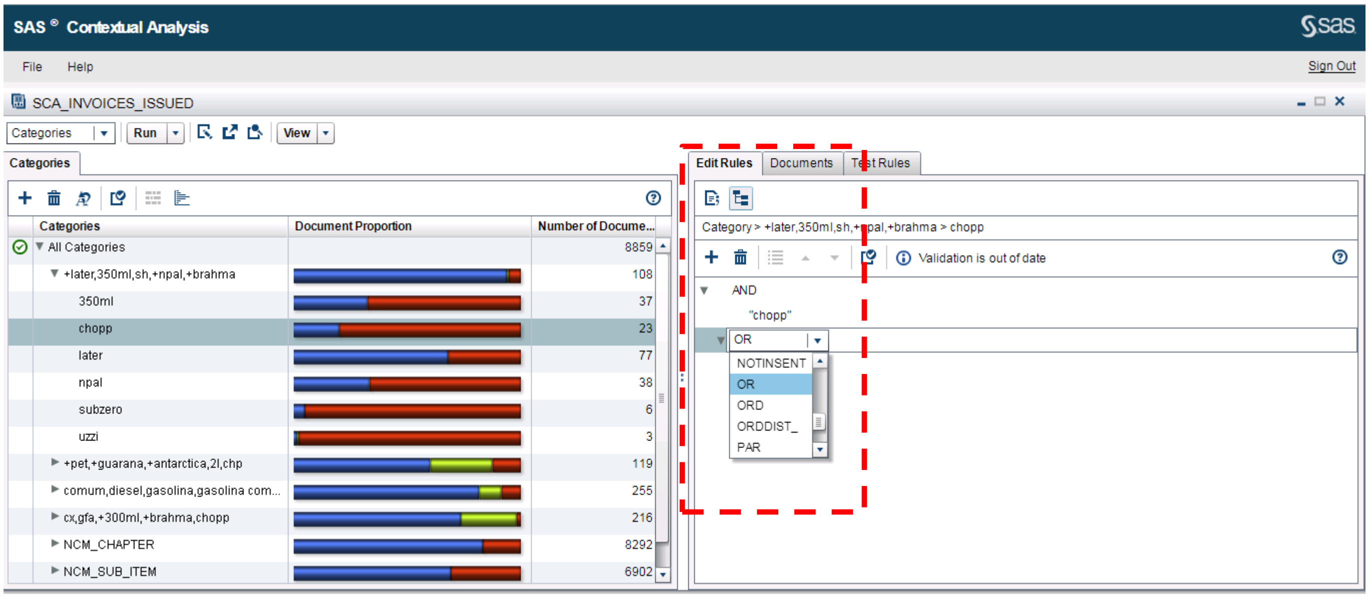
ID	Text
1300	BRAHMA CHOPP GFA VD 300ML CX C/23 31,6036
1301	BRAHMA CHOPP GFA VD 300ML CX C/23 25,3000
1303	BUDWEISER LATA 350ML SH C/12 NPAL 27,9600
1305	ANTARCTICA SUBZERO LATA 350ML SH C 12 NPAL 14,6724
1306	BRAHMA CHOPP LT 473ML SH C/12 NPAL 22,5888
1308	CERVEJA SKOL LATA 350ML
1309	BRAHMA CHOPP GFA VD 300ML CX C/23 29,9000
1310	KAISER PILSEN LATA 473ML 12X1FT
1311	BRAHMA CHOPP GFA VD 300ML CX C/23 31,1470
1312	BRAHMA CHOPP LT 473ML SH C/12 NPAL 26,7300
1313	BRAHMA CHOPP GFA VD 1L COM TTC 44,4240
1314	SKOL LT 269ML SH C15 NPAL 28,8000
1316	SKOL LATA 350ML SH C/12 NPAL 23,9467
1317	BRAHMA CHOPP 300ML 14,3200

Display 16. Examples of a Category and Its Taxonomies

CATEGORIZATION: EDIT RULES AND SCORE NEW DOCUMENTS

One of the first challenges for the business analyst is to develop a taxonomy that automatically categorizes invoice issues and that is updated in a recurring and more accurate manner according to the NCM. The results and benefits of accomplishing this are immediate, such as properly identifying the tax rate (for example, ICMS) and identifying possible anomalies in the application of the tax rate.

Display 17 shows the new category available in the Categories section. At this point, the analyst can improve the categorization process with the inclusion of his business knowledge on the **Edit Rules** tab.



Categories	Document Proportion	Number of Docume...
All Categories		8859
+later,350ml,sh,+npal,+brahma		108
350ml		37
chopp		23
later		77
npal		38
subzero		6
uzzi		3
+pet,+guarana,+antarctica,2l,chp		119
comum,diesel,gasolina,gasolina com...		255
cx,gfa,+300ml,+brahma,chopp		216
NCM_CHAPTER		8292
NCM_SUB_ITEM		6902

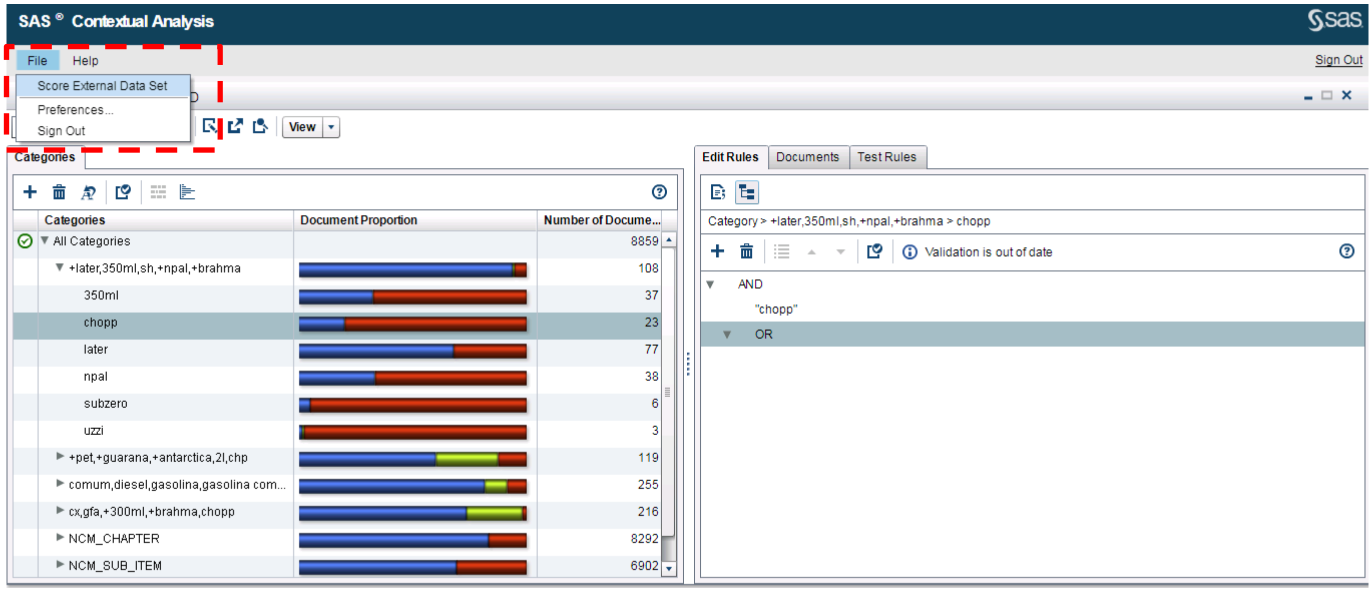
Category > +later,350ml,sh,+npal,+brahma > chopp

Validation is out of date

```
AND
    "chopp"
    OR
```

Display 17. Examples of Categories and Their Taxonomies

You can also use models developed in SAS Contextual Analysis to score additional text data. Select **File>Score External Data Set** (see Display 18). A window appears in which you can identify the textual data that you want to score. Additionally, you can view and export the DS2 macro code used to define concepts, sentiment, and categories for use anywhere you can run SAS.



Categories	Document Proportion	Number of Docume...
All Categories		8859
+later,350ml,sh,+npal,+brahma		108
350ml		37
chopp		23
later		77
npal		38
subzero		6
uzzi		3
+pet,+guarana,+antarctica,2l,chp		119
comum,diesel,gasolina,gasolina com...		255
cx,gfa,+300ml,+brahma,chopp		216
NCM_CHAPTER		8292
NCM_SUB_ITEM		6902

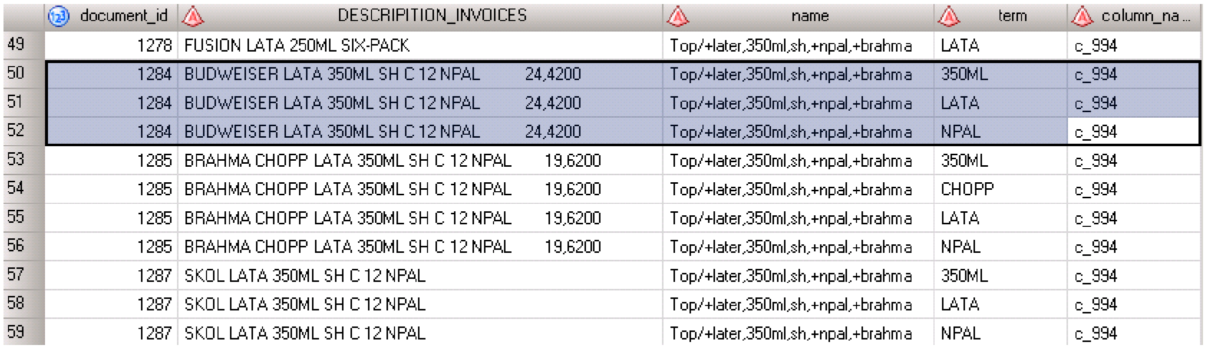
Category > +later,350ml,sh,+npal,+brahma > chopp

Validation is out of date

```
AND
    "chopp"
  OR
```

Display 18. Score External Data Set

Display 19 shows the results after categorization. The variable *document_id* is the ID of the invoices; the variable *name* is the name of the category, and the text with the description of the notes is in the Description_Invoices column.



	document_id	DESCRIPITION_INVOICES	name	term	column_na...
49	1278	FUSION LATA 250ML SIX-PACK	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
50	1284	BUDWEISER LATA 350ML SH C 12 NPAL 24,4200	Top/+later,350ml,sh,+npal,+brahma	350ML	c_994
51	1284	BUDWEISER LATA 350ML SH C 12 NPAL 24,4200	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
52	1284	BUDWEISER LATA 350ML SH C 12 NPAL 24,4200	Top/+later,350ml,sh,+npal,+brahma	NPAL	c_994
53	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	350ML	c_994
54	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	CHOPP	c_994
55	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
56	1285	BRAHMA CHOPP LATA 350ML SH C 12 NPAL 19,6200	Top/+later,350ml,sh,+npal,+brahma	NPAL	c_994
57	1287	SKOL LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	350ML	c_994
58	1287	SKOL LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	LATA	c_994
59	1287	SKOL LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	NPAL	c_994

Display 19. Categorization Result

INPUTS FOR CALCULATING THE FINAL WEIGHTED AVERAGE CONSUMER PRICE

The calculation of the final weighted consumer average price is updated frequently, and the values for some products rise more than others. In Brazil, the most common products for which the ICMS is calculated based on the final weighted average consumer price are fuels, drinks, and cosmetics, among other goods.

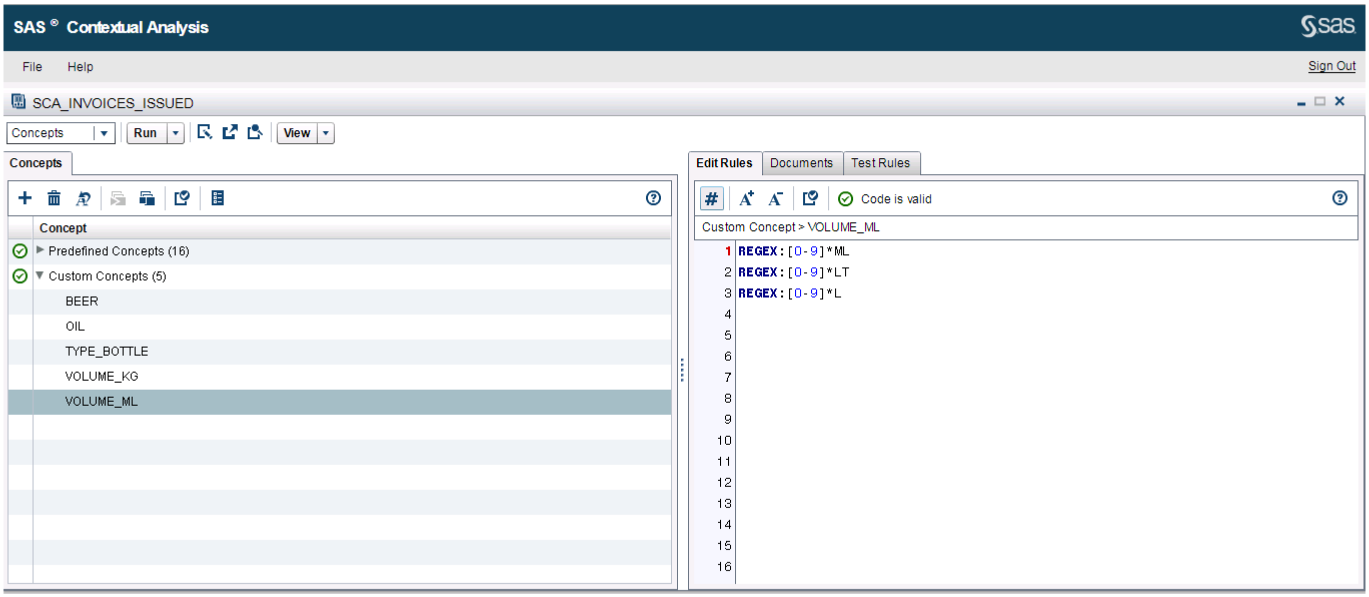
The taxpayer needs to be aware of this calculation and determine whether they are subject it. Otherwise, taxpayers might end up doing their ICMS calculations erroneously.

For this reason, there is a need to extract concepts like volume, type, quantity, and product name from the thousands or millions of Electronic Tax Invoices for inclusion in the calculation of the final weighted average consumer price.

HOW SAS CONTEXTUAL ANALYSIS ENRICHES THE CALCULATION

SAS Contextual Analysis uses language interpretation and text interpretation (LITI) syntax and its concept rules to recognize terms like. kg, ml, bottle, and so on, in context, so that you can extract only concepts in a document (for example, “Budweiser 355ML”) that match your rule.

In Display 20, you can see a custom concept node named VOLUME_LT and regular expressions (Regex syntax). These elements will extract all Electronic Tax Invoices in our data source that contain “LT” and all combinations that include numbers (RECEX: [0-9]*LT). The operator - is a wildcard that matches any character.



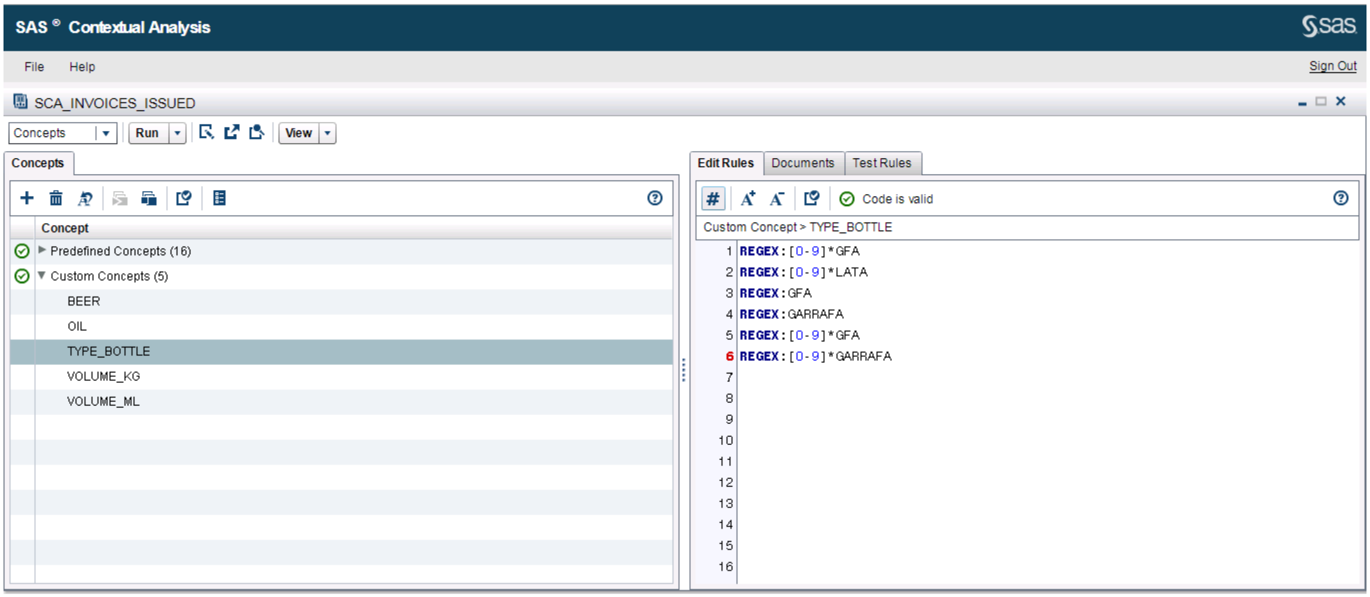
Concept
- Predefined Concepts (16)
- Custom Concepts (5)
  - BEER
  - OIL
  - TYPE_BOTTLE
  - VOLUME_KG
  - VOLUME_ML

Custom Concept > VOLUME_ML

```
REGEX:[0-9]*ML
REGEX:[0-9]*LT
REGEX:[0-9]*L
```

Display 20. Custom Concepts and Editing Rules for VOLUME_ML

Display 21 shows the rule for identifying all Electronic Tax Invoices for the concept node TYPE_BOTTLE that contain the terms "GFA, LATA, GARRAFA" and any number combination. Each document is evaluated separately for matches (shown in Display 22).

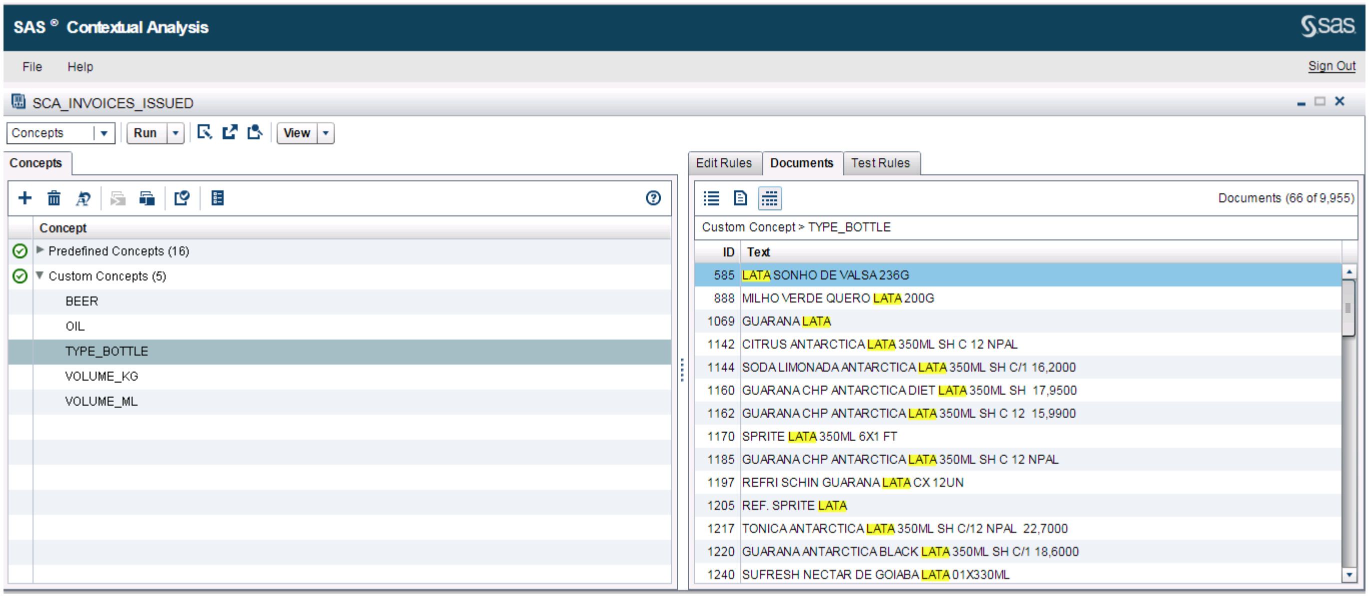


Concept
- Predefined Concepts (16)
- Custom Concepts (5)
  - BEER
  - OIL
  - TYPE_BOTTLE
  - VOLUME_KG
  - VOLUME_ML

Custom Concept > TYPE_BOTTLE

```
REGEX:[0-9]*GFA
REGEX:[0-9]*LATA
REGEX:GFA
REGEX:GARRAFA
REGEX:[0-9]*GFA
REGEX:[0-9]*GARRAFA
```

Display 21. Custom Concepts and Editing Rules for TYPE_BOTTLE



Concept
▶ Predefined Concepts (16)
▼ Custom Concepts (5)
BEER
OIL
TYPE_BOTTLE
VOLUME_KG
VOLUME_ML

Documents (66 of 9,955)

Custom Concept > TYPE_BOTTLE

ID	Text
585	LATA SONHO DE VALSA 236G
888	MILHO VERDE QUERO LATA 200G
1069	GUARANA LATA
1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL
1144	SODA LIMONADA ANTARCTICA LATA 350ML SH C/1 16,2000
1160	GUARANA CHP ANTARCTICA DIET LATA 350ML SH 17,9500
1162	GUARANA CHP ANTARCTICA LATA 350ML SH C 12 15,9900
1170	SPRITE LATA 350ML 6X1 FT
1185	GUARANA CHP ANTARCTICA LATA 350ML SH C 12 NPAL
1197	REFRI SCHIN GUARANA LATA CX 12UN
1205	REF. SPRITE LATA
1217	TONICA ANTARCTICA LATA 350ML SH C/12 NPAL 22,7000
1220	GUARANA ANTARCTICA BLACK LATA 350ML SH C/1 18,6000
1240	SUFRESH NECTAR DE GOIABA LATA 01X330ML

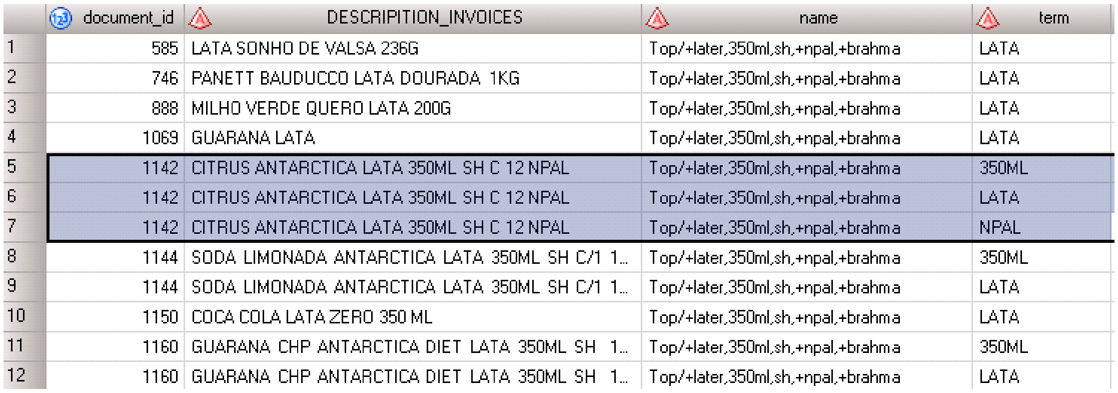
Display 22. Results of the Custom Rule for TYPE_BOTTLE

ANALYTICAL BASE TABLE FOR THE CALCULATION

Today, the final weighted average consumer price is typically obtained from sample surveys of final consumer prices. Such surveys can be ordered from the Finance Secretary.

Display 23 shows an example created in the SAS Contextual Analysis, which shows a possible analytical basis that can be used in the final weighted final consumer price calculation. The variable *document_id* represents the identification of the electronic invoice, DESCRIPITION_INVOICES contains the contents of the invoice, *name* is the category, and *term* is the result of extracting the electronic invoice concepts.

As an example, we could calculate the average final consumer price of all invoices classified as BEERS (*name*) and sold in cans of 355ML (*term* = "can" and name = "+ chopp + brhama + ...") . This process would already be automated, and it would be possible to generate reports in SAS Visual Analytics. This same logic would enrich the calculation for other items like food, building materials, and so on.



	document_id	DESCRIPITION_INVOICES	name	term
1	585	LATA SONHO DE VALSA 236G	Top/+later,350ml,sh,+npal,+brahma	LATA
2	746	PANETT BAUDUCCO LATA DOURADA 1KG	Top/+later,350ml,sh,+npal,+brahma	LATA
3	888	MILHO VERDE QUERO LATA 200G	Top/+later,350ml,sh,+npal,+brahma	LATA
4	1069	GUARANA LATA	Top/+later,350ml,sh,+npal,+brahma	LATA
5	1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	350ML
6	1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	LATA
7	1142	CITRUS ANTARCTICA LATA 350ML SH C 12 NPAL	Top/+later,350ml,sh,+npal,+brahma	NPAL
8	1144	SODA LIMONADA ANTARCTICA LATA 350ML SH C/1 1...	Top/+later,350ml,sh,+npal,+brahma	350ML
9	1144	SODA LIMONADA ANTARCTICA LATA 350ML SH C/1 1...	Top/+later,350ml,sh,+npal,+brahma	LATA
10	1150	COCA COLA LATA ZERO 350 ML	Top/+later,350ml,sh,+npal,+brahma	LATA
11	1160	GUARANA CHP ANTARCTICA DIET LATA 350ML SH 1...	Top/+later,350ml,sh,+npal,+brahma	350ML
12	1160	GUARANA CHP ANTARCTICA DIET LATA 350ML SH 1...	Top/+later,350ml,sh,+npal,+brahma	LATA

Display 23. Calculating the Final Weighted Average Consumer Price

CONCLUSION

This paper shows how you can use SAS Contextual Analysis to automate the process of product categorization and create custom concepts, using data that supports the calculation of the tax for the Electronic Tax Invoice. This methodology can be used in other Mercosul countries to reduce analysis time. This methodology can also improve governance, trust, and accuracy for the validation of invoice issues.

REFERENCES

COSTA, Leonardo De Andrade. 2015. “Processo Administrativo Tributário.” FGV Direito Rio. Available https://direitorio.fgv.br/sites/direitorio.fgv.br/files/u100/processo_administrativo_tributario_2015-2.pdf

SAS Contextual Analysis 14.1: Reference Help.

Website “O seu Portal Jurídico da internet-Âmbito Jurídico”. Available http://www.ambito-juridico.com.br/site/. Accessed on February 20, 2018.

Website “Portal da Nota Fiscal Electrônica”. Available http://www.nfe.fazenda.gov.br/portal/principal.aspx. Accessed on February 20, 2018.

ACKNOWLEDGMENTS

The author thanks Mauricio Fonseca Fernandino and Secretaria de Fazenda de Minas Gerais for Data Sources. The author thanks Joan Keyser and Amy Wolfe for help in reviewing the text.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Alexandre Carvalho
SAS Institute Brasil
55 21 99121-3280
Alexandre.carvalho@sas.com
https://br.linkedin.com/in/alexandrecarvalho

Paper SAS1986-2018

Using SAS® Text Analytics to Assess International Human Trafficking Patterns

Tom Sabo, Adam Pilz, SAS Institute Inc

ABSTRACT

The US Department of State (DOS) and other humanitarian agencies have a vested interest in assessing and preventing human trafficking in its many forms. A subdivision within the DOS releases publicly facing Trafficking in Persons (TIP) reports for approximately 200 countries annually. These reports are entirely freeform text, though there is a richness of structure hidden within the text. How can decision-makers quickly tap this information for patterns in international human trafficking?

This paper showcases a strategy of applying SAS® Text Analytics to explore the TIP reports and apply new layers of structured information. Specifically, we identify common themes across the reports, use topic analysis to identify a structural similarity across reports, identifying source and destination countries involved in trafficking, and use a rule-building approach to extract these relationships from freeform text. We subsequently depict these trafficking relationships across multiple countries in SAS® Visual Analytics, using a geographic network diagram that covers the types of trafficking as well as whether the countries involved are invested in addressing the problem. This ultimately provides decision-makers with big-picture information about how to best combat human trafficking internationally.

INTRODUCTION

Human trafficking is one of the most tragic human rights issues of our time. It splinters families, distorts global markets, undermines the rule of law, and spurs other transnational criminal activity. It threatens public safety and national security[1]. The International Labour Organization estimates that there are 20.9 million victims of human trafficking globally, and that forced labor and human trafficking generates 150 billion dollars in illicit gains annually. Of the 20.9 million victims, 26% are children, and 55% are women and girls[2].

The U.S. Department of state produces the Trafficking in Persons (TIP) report annually. It assesses the state of human trafficking in approximately 200 countries. This report is the U.S. Government's principal diplomatic tool to engage foreign governments on human trafficking. It is also the world's most comprehensive resource of governmental anti-trafficking efforts and reflects the U.S. Government's commitment to global leadership on this key human rights and law enforcement issue. It is used by the U.S. Government and worldwide as a tool to engage in dialogs to advance anti-trafficking reforms, and examine where resources are most needed. Freeing victims, preventing trafficking, and bringing traffickers to justice are the ultimate goals of the report and of the U.S Government's anti-trafficking policy[1]. However, the insights in these reports are scattered across hundreds of free-form text documents. How can we make the data in these reports more accessible to the broad audience that it supports, and how can we better envision patterns in the data which can be used to combat human trafficking?

This paper showcases a combination of SAS technology to identify patterns in the reports ultimately making the information more accessible to the various stakeholders. In particular, we will show how SAS can be used to identify links between source and destination countries, and visually depict these geospatial patterns in a network diagram. In this process, we will apply text analytics and visualization capabilities, primarily from SAS Visual Text Analytics and SAS Visual Analytics, available through SAS Viya. We will answer the following questions.

- Can we assess overall themes in international trafficking from the reports?
- Can we identify more focused patterns in trafficking, such as how women and children are seeking and achieving refuge?
- Can we identify and geospatially visualize patterns in trafficking across countries, including who is being trafficked (men, women, children), what type of trafficking is occurring (labor or sex trafficking), and whether the countries in question are in cooperation to address the problem?

By the end of this paper, the reader will learn how to apply the full SAS analytics lifecycle to this problem[3]. In this case, the analytics lifecycle includes data acquisition, unstructured and structured data management, text analytics, and network visualization. The reader will also gain an understanding of some key features available in SAS Visual Text Analytics, including similarity scores and fact extraction. We will also highlight some functionality common to the SAS text analytics products, including capabilities available across SAS Visual Text Analytics, SAS Contextual Analysis, and SAS Text Miner.

DATA ACQUISITION AND DATA MANAGEMENT

We obtained the data for each country narrative from the U.S. Department of State Trafficking in Persons report for 2017 using a script that accessed the following link: https://www.state.gov/j/tip/rls/tiprpt/countries/2017/index.htm. A slight modification to the script enabled us to obtain country narrative data from 2013-2016. Each report contains summary information about trafficking in the country, as well as several subsections. Subsections include recommendations, how the country prosecutes human traffickers, how the country protects victims, how the country prevents trafficking, and an overall trafficking profile.

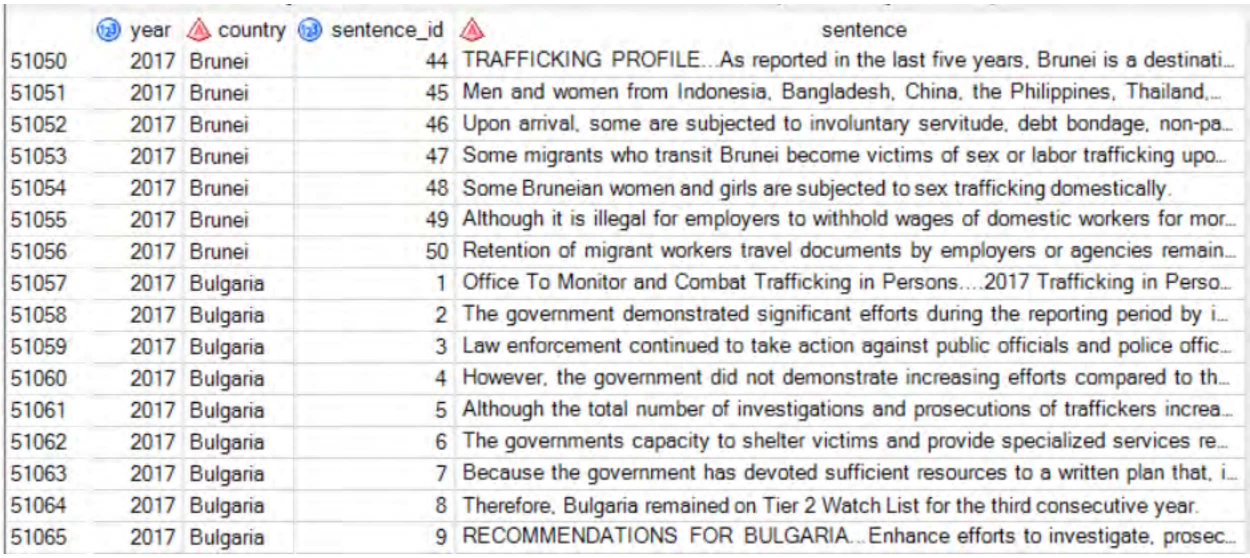
The country level trafficking reports are several pages in length. When working with documents greater than a page or two, it is helpful to apply some level of tokenization prior to text analytics[4]. Longer documents are more likely to have multiple themes embedded within. Tokenization breaks the documents up into smaller chunks, while maintaining a reference for each chunk to the larger documents. Then, patterns that appear across chunks can be readily surfaced using the capabilities at our disposal. This makes our algorithms more likely to identify these discrete themes or topics within documents.

For this effort, we applied sentence level tokenization. The following is SAS code we used for sentence level tokenization. In this case, it accepted as input a SAS data set that contained a number of rows of data, each containing a country level narrative from the TIP reports from 2013-2017.

```
/*Define the library where the data set is stored*/
libname _mylib 'D:\data\SamplePDF';

/*Define the data set for which you desire tokenized sentences*/
%let dsn = _mylib.output_sas_data;

/*Define the text variable to parse*/
%let text_var = text;

/*Strip the data and create an index*/
data temp (compress=yes); set &dsn;
      doc_id = _n_;
run;

/*parse the data set*/
proc hptmine data=temp;
      doc_id doc_id;
      var &text_var;
      parse
            nostemming notagging nonoungroups shownumpunct
            entities = none
            outpos   = position
            buildindex ;
      performance details ;
run;
```

```
proc sort data=position;
     by document sentence _start_;
run;

data sentenceSize (compress=yes);
     retain document start size;
     set position;
     by document sentence;
     if First.sentence then start=_start_+1;
     if Last.sentence then do;
          size=_end_ -start+2;
          output;
     end;
     keep document start size;
run;

/*Clean up*/
proc delete data=position; run;

data sentenceObs(compress=yes);
     length sentences $1000;
     merge sentenceSize(in=A ) temp (rename=(doc_id=document) );
     by document;
     if A then do;
          sentences=substrn(&text_var,start,size);
          output;
     end;
     keep sentences document;
run;

/*Clean up*/
proc delete data=sentenceSize; run;

data _mylib.output_sentences(compress=yes);
     set sentenceObs;
     by document;
     if first.document then sid = 1; else sid + 1;
run;

/*Clean up*/
proc delete data=sentenceObs; run;

/*view the data*/
proc print data=_mylib.output_sentences (obs=100);
run;
```

The output data set from the sentence tokenization procedure contained a row of data for each sentence in the original country level trafficking narratives, maintaining year, country, and sentence ID. This amounted to 63,648 rows of sentence level data. The following figure depicts a snapshot of the data. We took a 15,000 row sample of this sentence level data across all countries and years to use in the text analytics exercise described in the next section.



	year	country	sentence_id	sentence
51050	2017	Brunei	44	TRAFFICKING PROFILE...As reported in the last five years, Brunei is a destinati...
51051	2017	Brunei	45	Men and women from Indonesia, Bangladesh, China, the Philippines, Thailand,...
51052	2017	Brunei	46	Upon arrival, some are subjected to involuntary servitude, debt bondage, non-pa...
51053	2017	Brunei	47	Some migrants who transit Brunei become victims of sex or labor trafficking upo...
51054	2017	Brunei	48	Some Bruneian women and girls are subjected to sex trafficking domestically.
51055	2017	Brunei	49	Although it is illegal for employers to withhold wages of domestic workers for mor...
51056	2017	Brunei	50	Retention of migrant workers travel documents by employers or agencies remain...
51057	2017	Bulgaria	1	Office To Monitor and Combat Trafficking in Persons....2017 Trafficking in Perso...
51058	2017	Bulgaria	2	The government demonstrated significant efforts during the reporting period by i...
51059	2017	Bulgaria	3	Law enforcement continued to take action against public officials and police offic...
51060	2017	Bulgaria	4	However, the government did not demonstrate increasing efforts compared to th...
51061	2017	Bulgaria	5	Although the total number of investigations and prosecutions of traffickers increa...
51062	2017	Bulgaria	6	The governments capacity to shelter victims and provide specialized services re...
51063	2017	Bulgaria	7	Because the government has devoted sufficient resources to a written plan that, i...
51064	2017	Bulgaria	8	Therefore, Bulgaria remained on Tier 2 Watch List for the third consecutive year.
51065	2017	Bulgaria	9	RECOMMENDATIONS FOR BULGARIA...Enhance efforts to investigate, prosec...

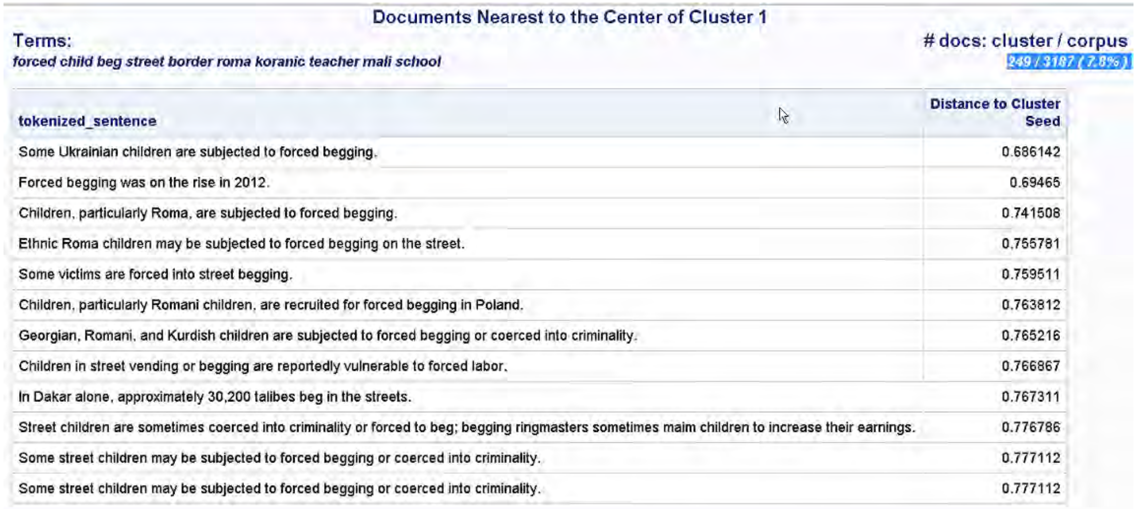
Figure 1: Sentence Level Country Narrative Data Used in Text Analytics

TEXT ANALYTICS

SAS now has a variety of capabilities in text analytics available in different solution packages. This includes capabilities in SAS Visual Text Analytics, available as a part of SAS Viya. This also includes capabilities present in SAS Text Miner, an add-on to SAS Enterprise Miner, and SAS Contextual Analysis, both available on any SAS 9 release. In this section on text analytics methods, we will show snapshots from individual solutions, and discuss which of the aforementioned SAS products also have the described capability.

IDENTIFYING OVERALL HUMAN TRAFFICKING TRENDS AND PATTERNS

One of the questions previously identified is whether we can assess overall themes in international trafficking from the reports. This is a capability available through an unsupervised machine learning method in text clustering. SAS assesses all the sentences across TIP reports and identifies key sets of terms that tend to occur together. For example, the terms “forced”, “child”, “beg”, and “street” tend to co-occur in the data along with other terms. These are indicative of a pattern across country narratives where children are coerced into begging. The following snapshot takes results from the text clustering capabilities of SAS Text Miner, and depicts the cluster results along with example sentences associated with the text cluster.



Documents Nearest to the Center of Cluster 1

Terms:
forced child beg street border roma koranic teacher mali school

docs: cluster / corpus
249 / 3187 (7.8%)

tokenized_sentence	Distance to Cluster Seed
Some Ukrainian children are subjected to forced begging.	0.686142
Forced begging was on the rise in 2012.	0.69465
Children, particularly Roma, are subjected to forced begging.	0.741508
Ethnic Roma children may be subjected to forced begging on the street.	0.755781
Some victims are forced into street begging.	0.759511
Children, particularly Romani children, are recruited for forced begging in Poland.	0.763812
Georgian, Romani, and Kurdish children are subjected to forced begging or coerced into criminality.	0.765216
Children in street vending or begging are reportedly vulnerable to forced labor.	0.766867
In Dakar alone, approximately 30,200 talibes beg in the streets.	0.767311
Street children are sometimes coerced into criminality or forced to beg; begging ringmasters sometimes maim children to increase their earnings.	0.776786
Some street children may be subjected to forced begging or coerced into criminality.	0.777112
Some street children may be subjected to forced begging or coerced into criminality.	0.777112

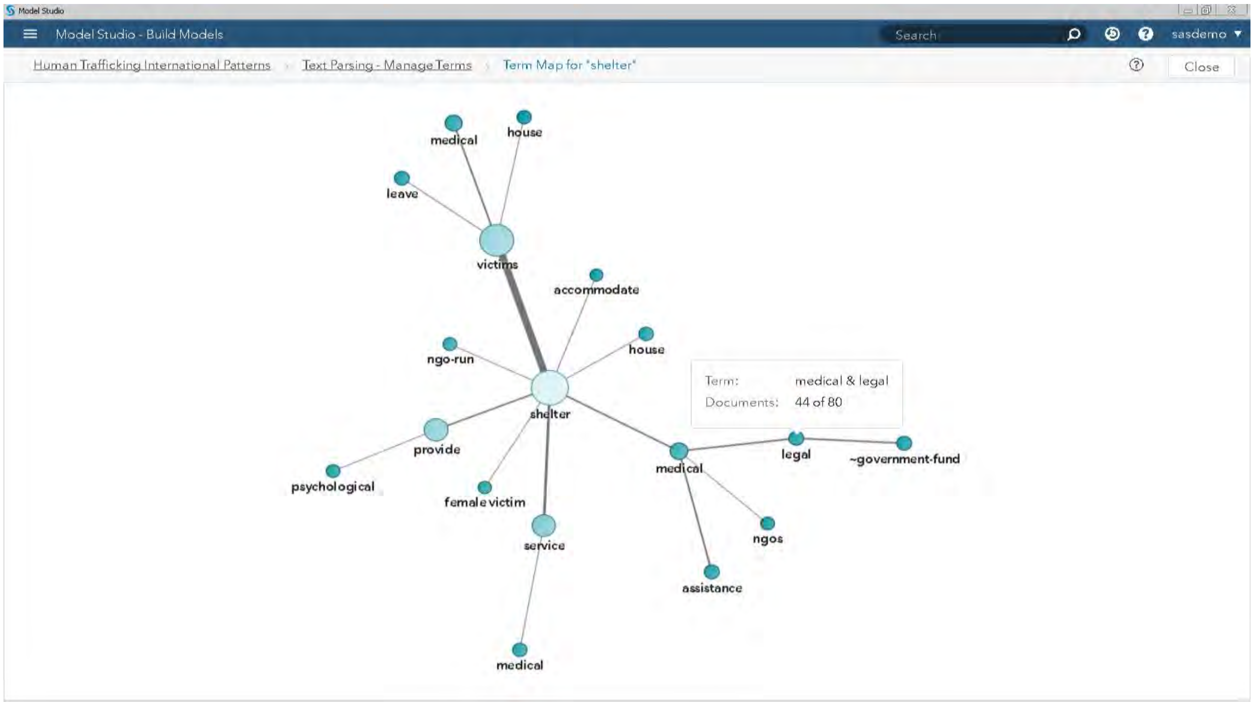
Figure 2: Themes Report Derived from SAS Text Miner Text Clustering

Similar results are available across all clusters and are indicative of a variety of themes in human trafficking. This includes where sex trafficking victims are typically exploited, groups who are subject to forced marriage and domestic servitude, what characteristics make individuals most vulnerable to human trafficking, and how debt bondage plays into human trafficking. Similar capabilities are available through the topics capability of both SAS Visual Text Analytics and SAS Contextual Analysis.

IDENTIFYING FOCUSED PATTERNS RELATED TO HUMAN TRAFFICKING

A second method to identify themes in the data is through a term map. In this method of interactive exploration, the user selects a term from the full list of extracted terms across all trafficking reports, and selects to view a term map of related terms and phrases. The user is then presented with a visual depiction of other terms and phrases that tend to be connected to the source term or phrase.

The example below depicts a term map surrounding the term “shelter”. This links other key terms and phrases, such as “provide” and “psychological”, indicating that shelters often provide psychological assistance. Another key linkage includes “female victim”, denoting who the shelters primarily serve. Finally, the term “medical” and “legal” tend to be associated with shelters, indicating other types of aid that are received at shelters. In the example below, in the 80 sentences across all reports that contain the term “shelter”, 44 of them also contain the terms “medical” and “legal”. This type of analysis provides quantitative data to advance anti-trafficking reforms, examine where resources are most needed, and can assist in determining where methods of providing assistance have been proven to be helpful. These methods can subsequently be implemented elsewhere.

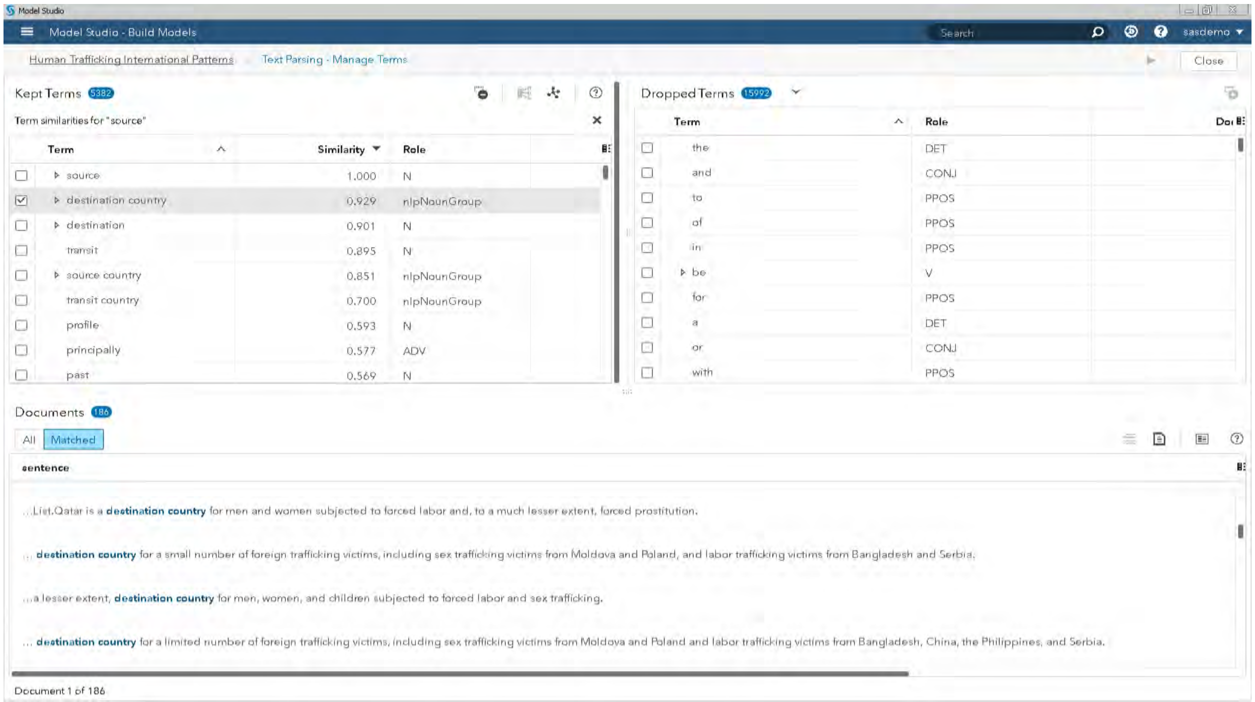


Term: medical & legal
Documents: 44 of 80

Figure 3: Term Map from SAS Visual Text Analytics Depicting Terms and Phrases Interconnected with the Term "Shelter"

Similar capability is available from the Text Filter node of SAS Text Miner, as well as from the Terms panel of SAS Contextual Analytics.

SAS Visual Text Analytics includes a unique capability across the SAS Text Analytics products that can identify term and phrase similarities to terms of interest. This differs from the term map capabilities in that algorithms identify terms used in a similar context to the selected term. From the Terms node of SAS Visual Text Analytics, the user can select a term, such as “source” in the example below, and view terms and phrases used in a similar context. In this visualization, SAS identifies terms including “source country”, “transit country”, and “destination country” used in a similar context, indicating that there are connective patterns in the text between countries that are a source of human trafficking victims, and countries where these victims become involved in human trafficking. The visualization also shows these terms in the context of the sentences in which they appear.


Model Studio
Model Studio - Build Models
Search
sasdemo
Human Trafficking International Patterns
Text Parsing - Manage Terms
Close
Kept Terms 5382
Dropped Terms 15992
Documents 186
All
Matched

Term similarities for "source"

	Term	Similarity	Role
□	source	1.000	N
☑	destination country	0.929	nlpNounGroup
□	destination	0.901	N
□	transit	0.895	N
□	source country	0.851	nlpNounGroup
□	transit country	0.700	nlpNounGroup
□	profile	0.593	N
□	principally	0.577	ADV
□	past	0.569	N

	Term	Role
□	the	DET
□	and	CONJ
□	to	PPOS
□	of	PPOS
□	in	PPOS
□	be	V
□	for	PPOS
□	a	DET
□	or	CONJ
□	with	PPOS

sentence

...List.Qatar is a **destination country** for men and women subjected to forced labor and, to a much lesser extent, forced prostitution.

... **destination country** for a small number of foreign trafficking victims, including sex trafficking victims from Moldova and Poland, and labor trafficking victims from Bangladesh and Serbia.

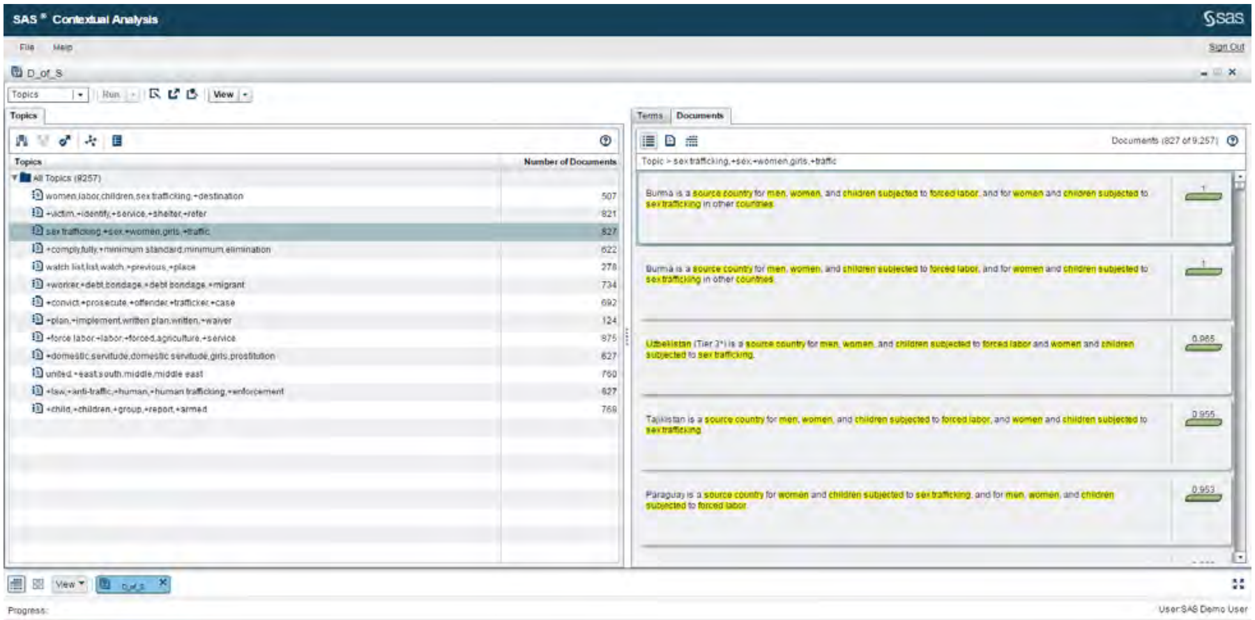
...a lesser extent, **destination country** for men, women, and children subjected to forced labor and sex trafficking.

... **destination country** for a limited number of foreign trafficking victims, including sex trafficking victims from Moldova and Poland and labor trafficking victims from Bangladesh, China, the Philippines, and Serbia.

Document 1 of 186

Figure 4: Visual Text Analytics Depiction of Term Similarity to the Term "source"

This connection between source, target, and transit countries is worth further exploration. To verify the depth of this pattern, we turn to the SAS Topics capability. In the example below taken from SAS Contextual Analysis, across 827 sentences, SAS identifies a theme (with no user input) between source countries and target countries. This theme also covers victims including men, women, and children, and the two forms of human trafficking, sex trafficking and labor trafficking.



Topics	Number of Documents
All Topics (9257)	
women,labor,children,sex trafficking,+destination	507
+victim,+identify,+service,+shelter,+refer	821
sex trafficking,+sex,+women,girls,+traffic	827
+comply,fully,+minimum standard,minimum,elimination	622
watch list,list,watch,+previous,+place	278
+worker,+debt bondage,+debt bondage,+migrant	734
+convict,+prosecute,+offender,+trafficker,+case	692
+plan,+implement,written plan,written,+waiver	124
+force labor,+labor,+forced,agriculture,+service	875
+domestic servitude,domestic servitude,girls,prostitution	627
united,+east,south,middle,middle east	760
+law,+anti-traffic,+human,+human trafficking,+enforcement	627
+child,+children,+group,+report,+armed	768

Documents (827 of 9,257)

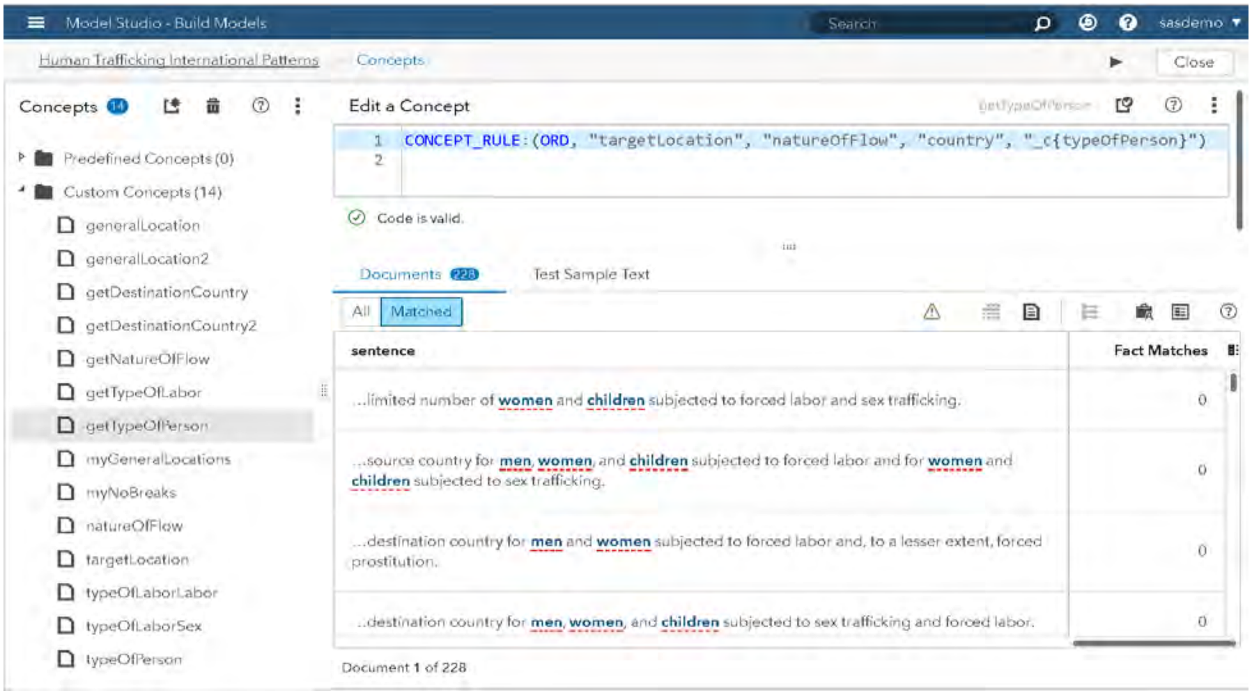
Topic > sex trafficking,+sex,+women,girls,+traffic

Burma is a source country for men, women, and children subjected to forced labor, and for women and children subjected to sex trafficking in other countries.	1
Burma is a source country for men, women, and children subjected to forced labor, and for women and children subjected to sex trafficking in other countries.	1
Uzbekistan (Tier 3*) is a source country for men, women, and children subjected to forced labor and women and children subjected to sex trafficking.	0.985
Tajikistan is a source country for men, women, and children subjected to forced labor, and women and children subjected to sex trafficking.	0.955
Paraguay is a source country for women and children subjected to sex trafficking, and for men, women, and children subjected to forced labor.	0.953

Figure 5: SAS Contextual Analysis Depicts a Topic Showing Network Connections in Human Trafficking

EXTRACTING PATTERNS IN HUMAN TRAFFICKING FOR NETWORK VISUALIZATION

Now that we have used exploratory text analytics capabilities to identify a pattern of interest, analysts might be interested in geospatially depicting the interconnection between countries on a world map over time. To prepare for this activity, it is necessary to develop rules to extract these patterns or facts via extraction rules. SAS Visual Text Analytics and SAS Contextual Analysis provide the capability to use a SAS proprietary rule-writing language called LITI to define parameters for fact extraction. Through the SAS Visual Text Analytics interface using LITI, we define rules to extract the victims of trafficking in context, including men, women, and children. This is depicted in the following example visualization of the rule editor and tester below.



```
CONCEPT_RULE:(ORD, "targetLocation", "natureOfFlow", "country", "_c{typeOfPerson}")
```

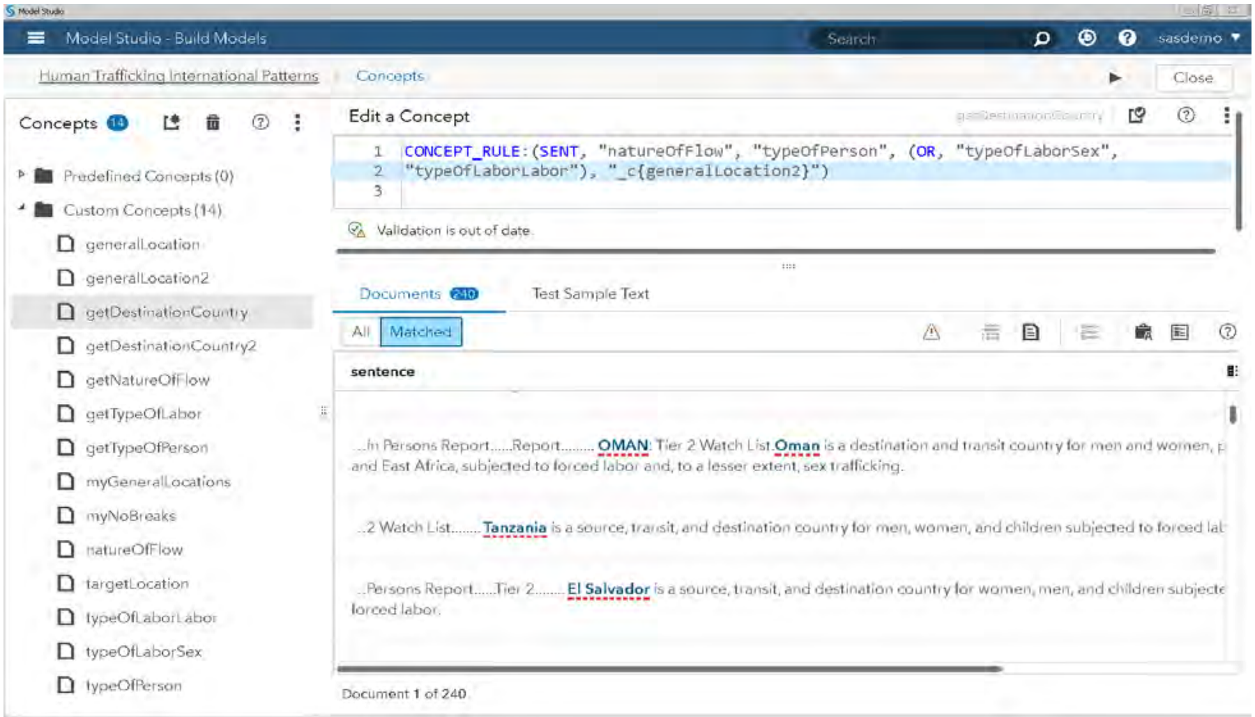
Code is valid.

sentence	Fact Matches
...limited number of **women** and **children** subjected to forced labor and sex trafficking.	0
...source country for **men**, **women**, and **children** subjected to forced labor and for **women** and **children** subjected to sex trafficking.	0
...destination country for **men** and **women** subjected to forced labor and, to a lesser extent, forced prostitution.	0
...destination country for **men**, **women**, and **children** subjected to sex trafficking and forced labor.	0

Document 1 of 228

Figure 6: LITI Rules in SAS Visual Text Analytics to Identify Victims of Human Trafficking

These definitions build upon themselves, and some rule definitions, such as a list of country names, become helper definitions when writing rules to extract a larger pattern. A set of rules, along with some post-processing, enables us to extract the full pattern of source countries, destination countries, Boolean indicators indicating the victims of trafficking and types of trafficking, and finally a cooperation indicator derived from the text for each pair of countries to determine whether they are working together to address the trafficking problem. The following screenshot depicts a rule which extracts destination countries for human trafficking victims.



```
CONCEPT_RULE:(SENT, "natureOfFlow", "typeOfPerson", (OR, "typeOfLaborSex",
"typeOfLaborLabor"), "_c{generalLocation2}")
```

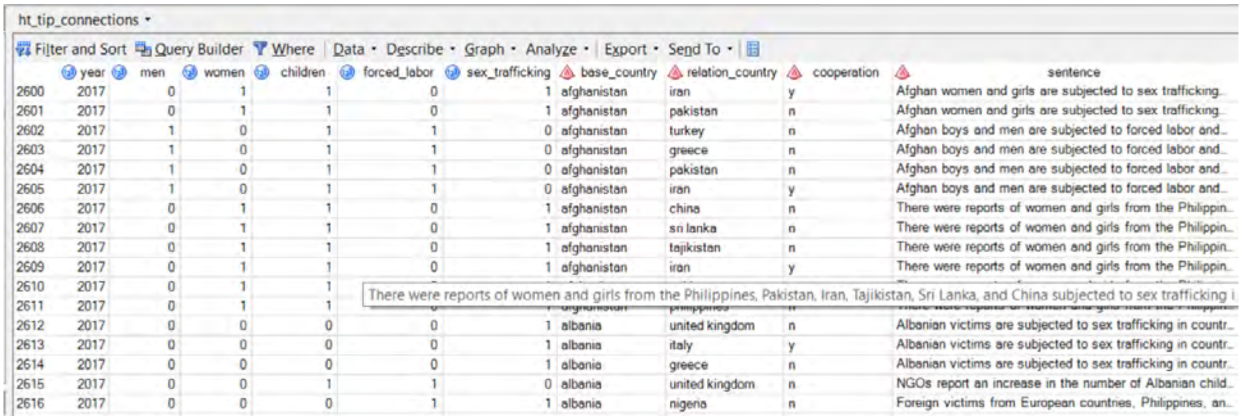
...In Persons Report......Report......... OMAN: Tier 2 Watch List.Oman is a destination and transit country for men and women, p
and East Africa, subjected to forced labor and, to a lesser extent, sex trafficking.

...2 Watch List........Tanzania is a source, transit, and destination country for men, women, and children subjected to forced lal

...Persons Report.....Tier 2........El Salvador is a source, transit, and destination country for women, men, and children subjecte
forced labor.

Figure 7: Concept Extraction in SAS Visual Text Analytics to Capture Human Trafficking Patterns

The purpose of SAS Visual Text Analytics is to develop and score these rules against source data to generate an additional data set used for visualization and interpretation of the data. In this case, we score the rules developed above to extract source/destination country patterns against the full 63,648 rows of sentence level data. In prior SAS Global Forum submissions, we've explored the output of a text analytics exercise in dashboard format, such as in assessing consumer financial compaints[5].These past use cases had the benefit of additional structured data in conjunction with the free-text field, such as a user complaint in context of structured geographical information, date of claim, type of claim, and whether a user who submitted the complaint received some form of monetary compensation. In this case, we develop a visualization using only structured data we generated from the unstructured reports, namely, the connection between the countries, including victim information, year of the report, type of trafficking, and cooperation indicator. Consider that we applied automated analysis to turn reams of text into visualization-ready structured data. Consider also that these processes could be immediately deployed for new sets of these reports as they emerge in 2018, 2019 and beyond! A snapshot of this data generated from Visual Text Analytics after postprocessing is depicted below.



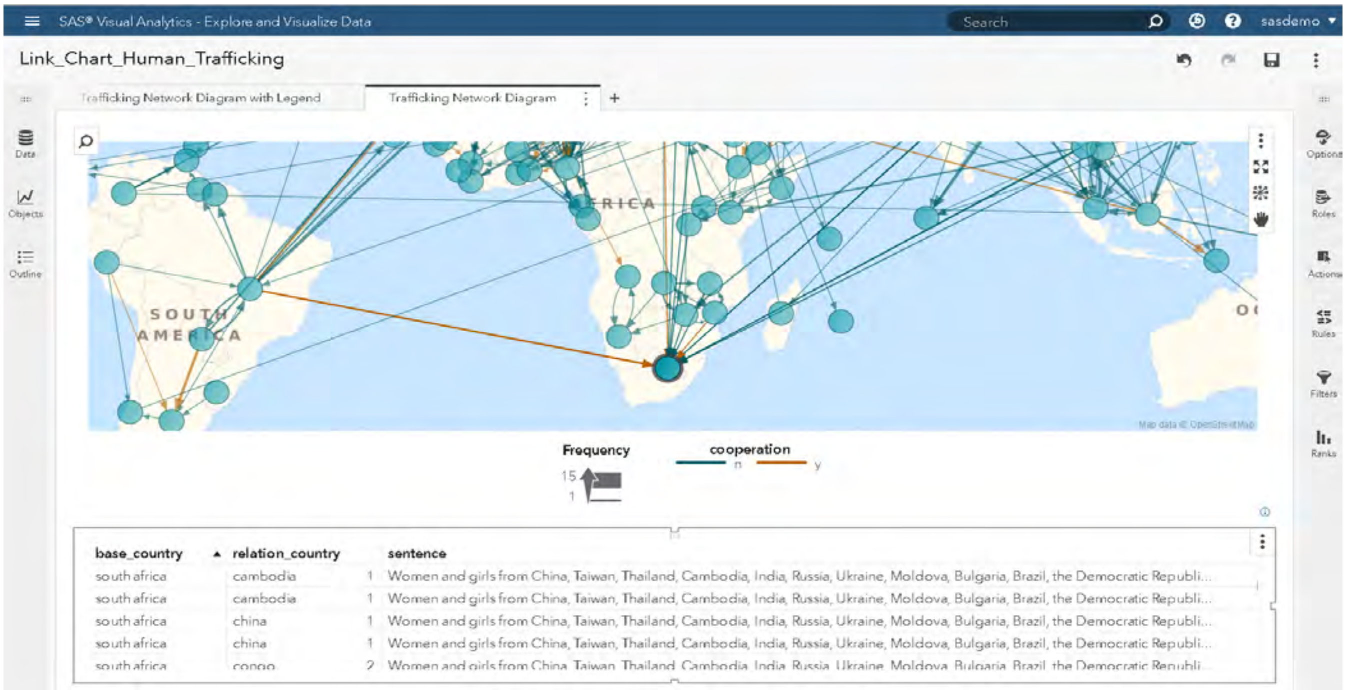
	year	men	women	children	forced_labor	sex_trafficking	base_country	relation_country	cooperation	sentence
2600	2017	0	1	1	0	1	afghanistan	iran	y	Afghan women and girls are subjected to sex trafficking...
2601	2017	0	1	1	0	1	afghanistan	pakistan	n	Afghan women and girls are subjected to sex trafficking...
2602	2017	1	0	1	1	0	afghanistan	turkey	n	Afghan boys and men are subjected to forced labor and...
2603	2017	1	0	1	1	0	afghanistan	greece	n	Afghan boys and men are subjected to forced labor and...
2604	2017	1	0	1	1	0	afghanistan	pakistan	n	Afghan boys and men are subjected to forced labor and...
2605	2017	1	0	1	1	0	afghanistan	iran	y	Afghan boys and men are subjected to forced labor and...
2606	2017	0	1	1	0	1	afghanistan	china	n	There were reports of women and girls from the Philippin...
2607	2017	0	1	1	0	1	afghanistan	sri lanka	n	There were reports of women and girls from the Philippin...
2608	2017	0	1	1	0	1	afghanistan	tajikistan	n	There were reports of women and girls from the Philippin...
2609	2017	0	1	1	0	1	afghanistan	iran	y	There were reports of women and girls from the Philippin...
2610	2017	0	1	1						
2611	2017	0	1	1	0	1	afghanistan	philippines	n	There were reports of women and girls from the Philippin...
2612	2017	0	0	0	0	1	albania	united kingdom	n	Albanian victims are subjected to sex trafficking in countr...
2613	2017	0	0	0	0	1	albania	italy	y	Albanian victims are subjected to sex trafficking in countr...
2614	2017	0	0	0	0	1	albania	greece	n	Albanian victims are subjected to sex trafficking in countr...
2615	2017	0	0	1	1	0	albania	united kingdom	n	NGOs report an increase in the number of Albanian child...
2616	2017	0	0	0	1	1	albania	nigeria	n	Foreign victims from European countries, Philippines, an...

There were reports of women and girls from the Philippines, Pakistan, Iran, Tajikistan, Sri Lanka, and China subjected to sex trafficking i

Figure 8: Visualization-Ready Data Generated by Scoring Rules from SAS Visual Text Analytics

NETWORK VISUALIZATION

We load the data generated from the text analytics exercise into SAS Visual Analytics. The following visualizations were accomplished on SAS Viya, but similar visualizations are available on SAS 9. Once the data is loaded and the option to create a new report is selected with that data, we select the Network Analysis object for our geospatial visualization. We select the option under “Network Display” to enable a Map background, which leverages OpenStreetMap by default. We convert the base country and relation country from a categorical data variable to a geography data variable based on the country name. These are set as source and target “roles” for the Network Analysis object. The link width is set to the frequency of connections between source and target countries, enabling thicker lines for relationships that span multiple years. The link color is set to the cooperation_indicator, highlighting links that involve cooperation between source and target countries in orange. Finally, the directionality of the links is assigned under the Link Direction option of the “Network Display” to “Source”, to show the links from source country to destination countries. The resulting diagram, initially centered around South Africa, is shown below.

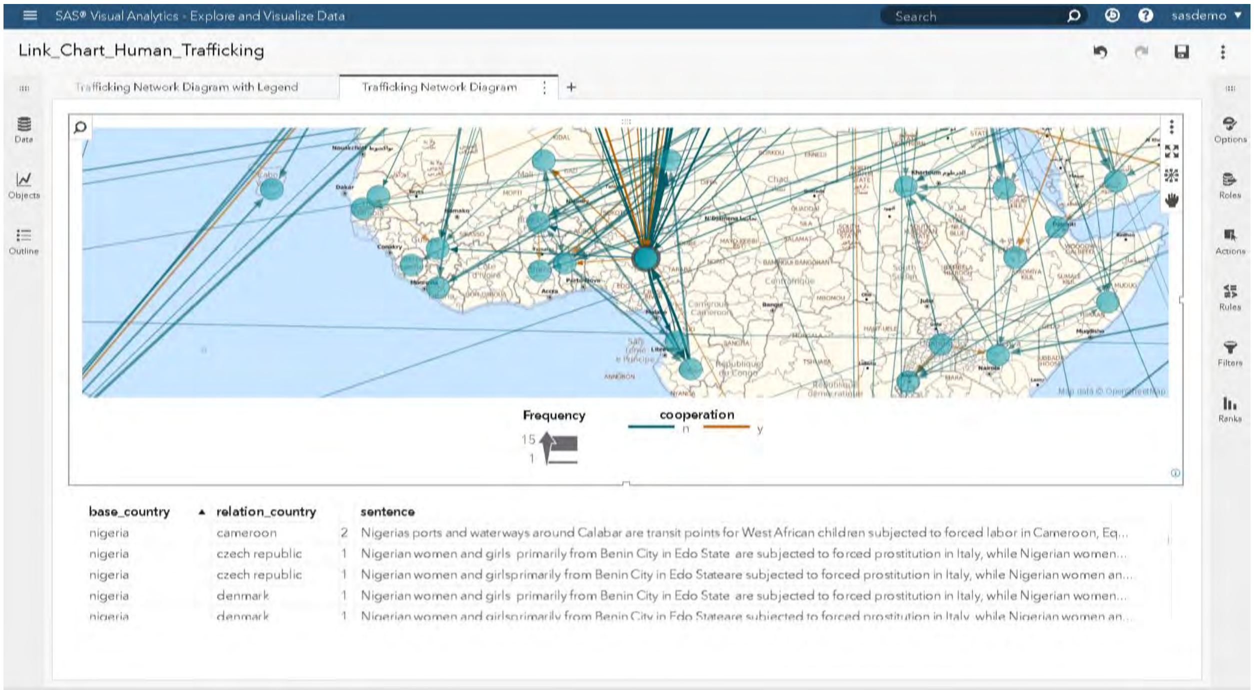

SAS® Visual Analytics - Explore and Visualize Data
Search
sasdemo
Link_Chart_Human_Trafficking
Trafficking Network Diagram with Legend
Trafficking Network Diagram
Data
Objects
Outline
Options
Roles
Actions
Rules
Filters
Ranks
AFRICA
SOUTH AMERICA
Map data © OpenStreetMap
Frequency
15
1
cooperation
n
y

base_country	▲ relation_country		sentence
south africa	cambodia	1	Women and girls from China, Taiwan, Thailand, Cambodia, India, Russia, Ukraine, Moldova, Bulgaria, Brazil, the Democratic Republi...
south africa	cambodia	1	Women and girls from China, Taiwan, Thailand, Cambodia, India, Russia, Ukraine, Moldova, Bulgaria, Brazil, the Democratic Republi...
south africa	china	1	Women and girls from China, Taiwan, Thailand, Cambodia, India, Russia, Ukraine, Moldova, Bulgaria, Brazil, the Democratic Republi...
south africa	china	1	Women and girls from China, Taiwan, Thailand, Cambodia, India, Russia, Ukraine, Moldova, Bulgaria, Brazil, the Democratic Republi...
south africa	congo	2	Women and girls from China, Taiwan, Thailand, Cambodia, India, Russia, Ukraine, Moldova, Bulgaria, Brazil, the Democratic Republi...

Figure 9: Network Analysis Diagram Showing Patterns of Trafficking in the Southern Hemisphere

This visualization displays the interconnection between all countries across the TIP reports from 2013-2017. It highlights groups of countries involved in trafficking with each other, such as the various countries in the south of Africa as well as South America. It also highlights countries that serve as hubs for larger international trafficking patterns. For each node selected, SAS Visual Analytics displays the text associated with those connections. In this case, it highlights lines from the TIP reports identifying victims of human trafficking in South Africa from source countries including China, Taiwan, Thailand, Cambodia, India, Russia, and Brazil. From here, the text can be assessed to verify the authenticity of the links. Some links, including the link between Brazil and South Africa, are depicted in orange. This shows that SAS identified a relationship in the text between those two countries indicating that they were working together to address the trafficking problem.

Connections between Nigeria and other African countries, as well as to countries in Europe and Asia are particularly strong as shown in the diagram below. This might warrant an analysis of other circumstantial evidence surrounding Nigeria, and we will explore this further in the discussion section of this paper.


SAS® Visual Analytics - Explore and Visualize Data
Search
sasdemo
Link_Chart_Human_Trafficking
Trafficking Network Diagram with Legend
Trafficking Network Diagram
Data
Objects
Outline
Options
Roles
Actions
Rules
Filters
Ranks
Frequency
15
1
cooperation
n
y
Map data © OpenStreetMap

base_country	▲ relation_country		sentence
nigeria	cameroon	2	Nigerias ports and waterways around Calabar are transit points for West African children subjected to forced labor in Cameroon, Eq...
nigeria	czech republic	1	Nigerian women and girls primarily from Benin City in Edo State are subjected to forced prostitution in Italy, while Nigerian women...
nigeria	czech republic	1	Nigerian women and girlsprimarily from Benin City in Edo Stateare subjected to forced prostitution in Italy, while Nigerian women an...
nigeria	denmark	1	Nigerian women and girls primarily from Benin City in Edo State are subjected to forced prostitution in Italy, while Nigerian women...
nigeria	denmark	1	Nigerian women and girlsprimarily from Benin City in Edo Stateare subjected to forced prostitution in Italy, while Nigerian women an...

Figure 10: Network Analysis Diagram Highlighting Patterns of Trafficking Surrounding Nigeria

Filters can be applied that showcase certain aspects of trafficking, such as labor trafficking only, or sex trafficking only. In the visualization below, only the patterns of trafficking extracted from the TIP reports that mention children are shown.

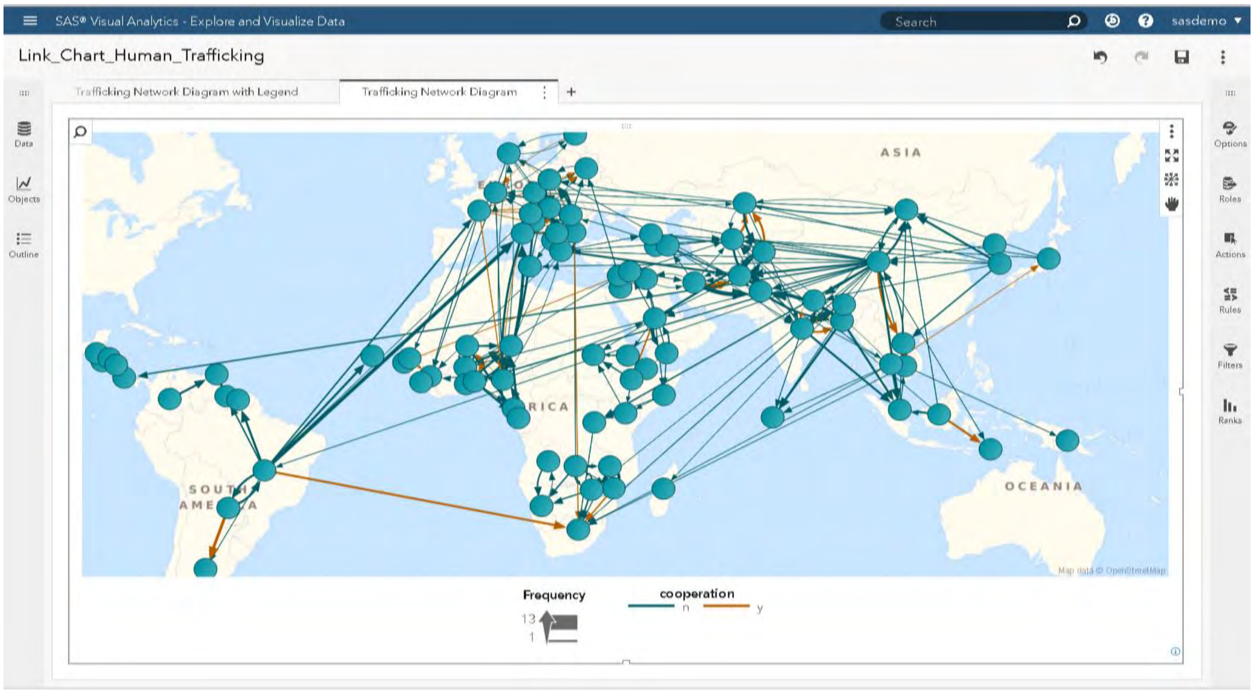
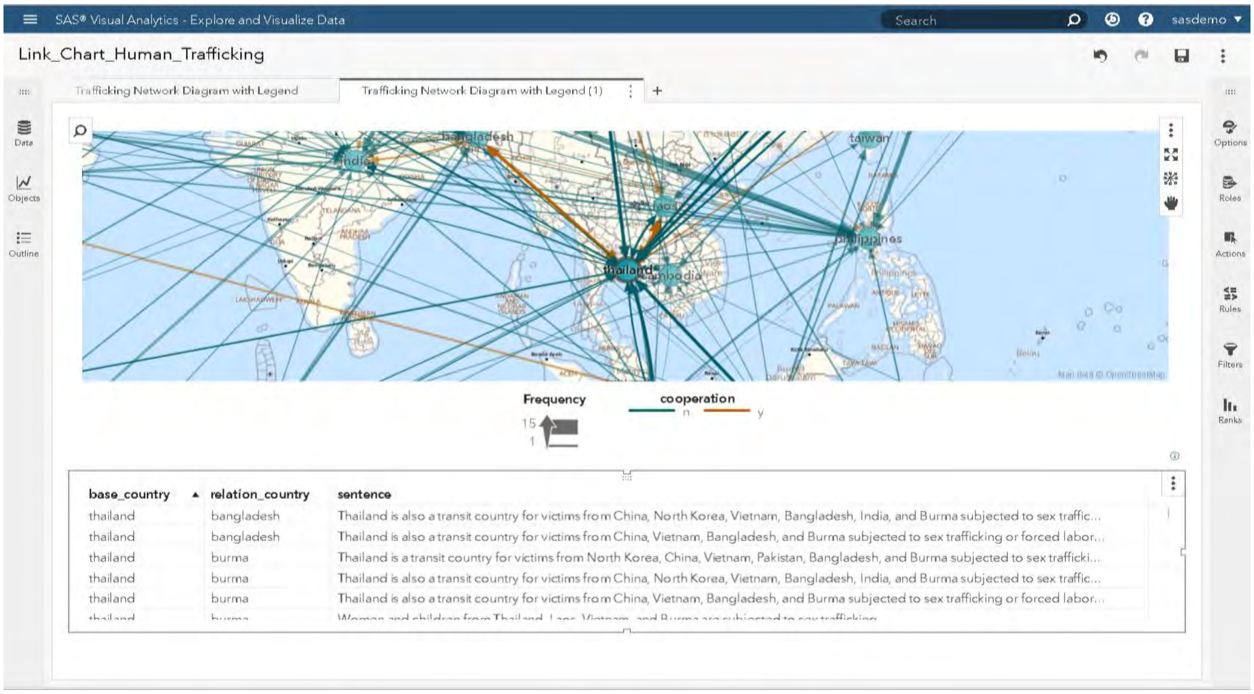


Figure 11: Network Analysis Diagram Depicting International Patterns of Child Trafficking

Finally, in considering visualization and interconnectedness between countries, the single links available in the TIP reports play into a much broader picture. Reports might mention connections such as “Nigeria is a source country for trafficking in other countries including…”. These single node-to-node connections become much more insightful when seen in the context of all the other node-to-node connections. This is particularly illuminating when countries are specifically called out as transit countries, and these connections in turn reveal second-degree and third-degree connections between source and destination countries. The following visualization reveals Thailand cited as a transit country for a variety of source and destination countries, revealing a larger pattern of international human trafficking.


SAS® Visual Analytics - Explore and Visualize Data
Search
sasdemo
Link_Chart_Human_Trafficking
Trafficking Network Diagram with Legend
Trafficking Network Diagram with Legend (1)
Data
Objects
Outline
Options
Roles
Actions
Rules
Filters
Ranks
bangladesh
india
taiwan
laos
philippines
thailand
cambodia
Frequency
15
1
cooperation
n
y

base_country ▲	relation_country	sentence
thailand	bangladesh	Thailand is also a transit country for victims from China, North Korea, Vietnam, Bangladesh, India, and Burma subjected to sex traffic...
thailand	bangladesh	Thailand is also a transit country for victims from China, Vietnam, Bangladesh, and Burma subjected to sex trafficking or forced labor...
thailand	burma	Thailand is a transit country for victims from North Korea, China, Vietnam, Pakistan, Bangladesh, and Burma subjected to sex trafficki...
thailand	burma	Thailand is also a transit country for victims from China, North Korea, Vietnam, Bangladesh, India, and Burma subjected to sex traffic...
thailand	burma	Thailand is also a transit country for victims from China, Vietnam, Bangladesh, and Burma subjected to sex trafficking or forced labor...
thailand	burma	Women and children from Thailand, Laos, Vietnam, and Burma are subjected to sex trafficking

Figure 12: Network Analysis Diagram Depicting Thailand as a Transit Country for International Trafficking

CONCLUSION AND DISCUSSION

In this paper, we showed how SAS could be used to obtain TIP reports from the US Department of State, identify patterns across those reports, and visually depict those patterns. We used the text analysis and visualization capabilities of SAS to answer three questions. First, we identified general trends in the TIP reports. Second, we identified focused themes, including themes around victims seeking shelter internationally. Finally, we extracted a geospatial pattern across all trafficking reports between source and target countries. We then depicted this visually in a network analysis diagram. The network analysis diagram included controls for filtering on trafficking victims, trafficking type, and the year of the report. These results enhance the ability of the U.S. Government and foreign nations worldwide to engage in dialogs advancing anti-trafficking reforms, and to examine where resources are most needed.

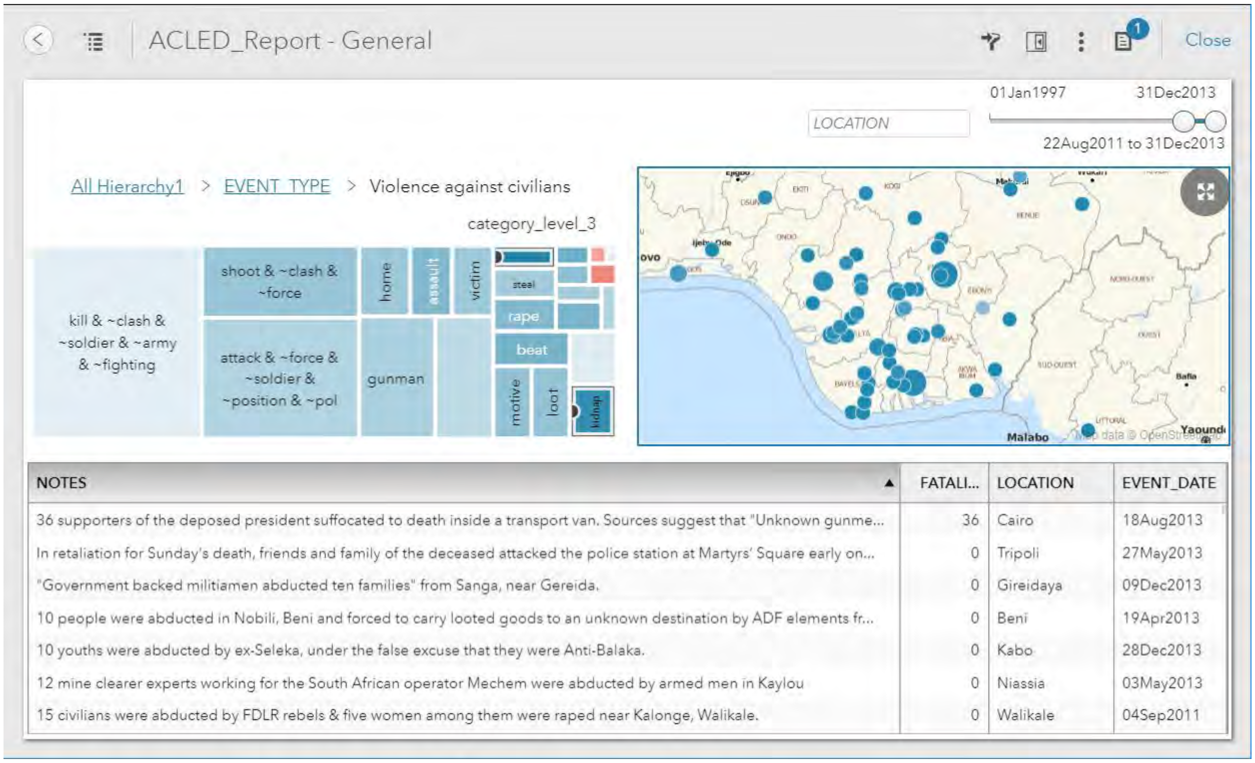
The analysis effort to identify the source and destination countries took approximately three days of dedicated effort. Contrast this with a manual effort to extract the same information from the reports. If we approximate 30 minutes per report to identify all the relevant connections that occurred in the data, with approximately 1000 reports, this would require 3 months of effort, or 30 times the time investment. Also, consider that the automated rules can score reports in upcoming years for connections, including 2018, 2019 and beyond at little extra time investment.

Analysis relies on the quality of the underlying data, and all analysis is fraught with challenges involving precision and recall. Precision in this case involves extracting only correct links, including getting the directionality of the connection correct. Recall involves extracting all of the links. Precision, in this analysis, can be improved by developing additional rules to ensure directionality accuracy in the links. Recall in this data set was influenced by factors including generality of information in the TIP reports. For example, the United States does not feature in any of the network links, as the United States is discussed in general terms in the reports, indicating that the United States is a source country and destination country for trafficking with a variety of foreign nations. This means that SAS is unable to extract a specific pattern related to the United States since specific countries it is connected to are not called out in the reports directly. Such insight provides additional feedback to the analysts developing these reports in terms of where specific patterns need to be built out upon the more general patterns. Such work can enhance understanding of the international patterns between several degrees of source and destination countries.

There are several different trafficking-related use cases, including drug trafficking and weapons trafficking. These tend to tie together with the human trafficking element. Other organizations who could potentially benefit from a trafficking solution include federal, state, and local law enforcement agencies. A solution that assesses and prioritizes trafficking-related leads could be set up from a law enforcement perspective, but could also address victim assistance. Regarding data that would assist law enforcement, search engines for classified ads become a repository of data that plays into human trafficking, particularly sex trafficking. They can be assessed to identify geographically where recruitment ads are spiking, where there are similar or emerging patterns in ads, and can ultimately assist law enforcement in identifying networks of trafficking-based organizations. There is a trafficking related pattern to data from financial organizations as well, including the major banks. The Financial Crimes Enforcement Network (FinCEN) has published guidelines to banks on recognizing activity that might be associated with human smuggling and human trafficking[6].

As mentioned, there are different sources of data that support the use case to assess patterns of international human trafficking. For example, to identify why Nigeria has a number of trafficking connections to a variety of countries in Africa, Europe, and Asia, we can examine data sources such as the Armed Conflict Location and Event Data project[7] (ACLED) to look for connections. In addition, we can apply machine learning and auto-categorization to the ACLED data as prescribed in a previous SAS Global Forum paper published in 2016[8].

In the screenshot below, we used a categorical hierarchy developed with machine learning against the ACLED data to explore themes in violence against civilians in Nigeria and the surrounding regions. The visualization depicts specific recorded instances of abduction and kidnaping, and drills down to the event text describing what happened. There is significant event traffic in Nigeria, depicting a destabilizing force that contributes to the vulnerability of its citizens to human trafficking. This analysis adds to the current evidence that many Nigerians seek work abroad due to extreme poverty, and are subsequently exploited for forced labor and prostitution. The data available from the TIP reports and the ACLED project is further reinforced by an exposé by CNN, where individuals who have sought work abroad as migrants from Niger and Nigeria among other African countries are sold at a slave auction[9].



NOTES	FATALI...	LOCATION	EVENT_DATE
36 supporters of the deposed president suffocated to death inside a transport van. Sources suggest that "Unknown gunme...	36	Cairo	18Aug2013
In retaliation for Sunday's death, friends and family of the deceased attacked the police station at Martyrs' Square early on...	0	Tripoli	27May2013
"Government backed militiamen abducted ten families" from Sanga, near Gereida.	0	Gireidaya	09Dec2013
10 people were abducted in Nobili, Beni and forced to carry looted goods to an unknown destination by ADF elements fr...	0	Beni	19Apr2013
10 youths were abducted by ex-Seleka, under the false excuse that they were Anti-Balaka.	0	Kabo	28Dec2013
12 mine clearer experts working for the South African operator Mechem were abducted by armed men in Kaylou	0	Niassia	03May2013
15 civilians were abducted by FDLR rebels & five women among them were raped near Kalonge, Walikale.	0	Walikale	04Sep2011

Figure 13: Visualization Depicting Kidnaping and Abduction Events in Nigeria and the Surrounding Countries Using Data from the ACLED Project

In summary, the analytics and visualizations presented here are an effort to show how data related to human trafficking can be transformed into actionable information. By taking advantage of data and analytics, data scientists and researchers are able to shine light on the problem, and thereby help international government, law enforcement, and victims advocacy groups find better ways to address it[10].

REFERENCES

1. U.S. Department of State. 2017. "Trafficking in Persons Report." Accessed February 2, 2018. https://www.state.gov/j/tip/rls/tiprpt/.

2. The Polaris Project. 2018. "The Facts." Accessed February 2, 2018. https://polarisproject.org/human-trafficking/facts.

3. Figallo-Monge, Manuel. "Pedal-to-the-Metal Analytics with SAS® Studio, SAS® Visual Analytics, SAS® Visual Statistics, and SAS® Contextual Analysis" Proceedings of the SAS Global Forum 2016 Conference. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings16/SAS6560-2016.pdf.

4. Albright, Russ. Cox, James. Jin, Ning. 2016. "Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations" Proceedings of the SAS Global Forum 2016 Conference. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6241-2016.pdf.

5. Sabo, Tom. 2017. "Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaints." *Proceedings of the SAS Global Forum 2017 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings17/SAS0282-2017.pdf.

6. United States Department of the Treasury, Financial Crimes Enforcement Network. 2014. “Advisory Information”. Accessed February 8, 2018. https://www.fincen.gov/resources/advisories/fincen-advisory-fin-2014-a008.

7. ACLED Data; Bringing Clarity to Crisis. 2018. “About”. Accessed February 12, 2018. http://www.acleddata.com/.

8. Sabo, Tom. 2016. “Extending the Armed Conflict Location and Event Data Project with SAS® Text Analytics.” Proceedings of the SAS Global Forum 2016 Conference. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6380-2016.pdf.

9. CNN. 2017. “Migrants being sold as slaves.” Accessed February 12, 2018. http://www.cnn.com/videos/world/2017/11/13/libya-migrant-slave-auction-lon-orig-md-ejk.cnn.

10. SAS. 2017. “Analytics tackles the scourge of human trafficking.” Accessed February 12, 2018. https://www.sas.com/en_us/insights/articles/analytics/analytics-tackles-human-trafficking.html.

ACKNOWLEDGMENTS

Thanks to Emily McRae and John Dillman for assisting with the visualizations, and thanks also to Emily for her review of this paper. Thanks to Mary Beth Ainsworth for her insight into the human trafficking problem nationally and internationally. Also, thanks to Dr. James R. Van Scotter for providing awareness into how local and state law enforcement can use data to identify human trafficking trends within the U.S.

RECOMMENDED READING

- Sabo, Tom. 2014. SAS Institute white paper. *“Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data.”* Available http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.
- Chakraborty, G., M. Pagolu, S. Garla. 2013. *Text Mining and Analysis; Practical Methods, Examples, and Case Studies Using SAS®. SAS Institute Inc.*

- Sabo, Tom. 2014. *"Uncovering Trends in Research Using Text Analytics with Examples from Nanotechnology and Aerospace Engineering." Proceedings of the SAS Global Forum 2014 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS061-2014.pdf
- Sabo, Tom. 2015. *"Show Me the Money! Text Analytics for Decision-Making in Government Spending." Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1661-2015.pdf.
- Reamy, Tom. 2016. Deep Text; Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data. Medford NJ: Information Today, Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Tom Sabo, Principal Solutions Architect
1530 Wilson Blvd.
Arlington, VA 22209
SAS Federal LLC
+1 (703) 310-5717
tom.sabo@sas.com
@mrTomSab

Adam Pilz, Senior Solutions Architect
121 W Trade St.
Charlotte, NC 28202
SAS Institute Inc
+1 (919) 348-6039
adam.pilz@sas.com

Paper SAS634-2017

An Efficient Way to Deploy and Run Text Analytics Models in Hadoop

Seung Lee, Xu Yang, and Saratendu Sethi, SAS Institute Inc.

ABSTRACT

Significant growth of the Internet has created an enormous volume of unstructured text data. In recent years, the amount of this type of data that is available for analysis has exploded. While the amount of textual data is increasing rapidly, an ability to obtain key pieces of information from such data in a fast, flexible, and efficient way is still posing challenges. This paper introduces SAS® Contextual Analysis In-Database Scoring for Hadoop, which integrates SAS® Contextual Analysis with the SAS® Embedded Process. SAS® Contextual Analysis enables users to customize their text analytics models in order to realize the value of their text-based data. The SAS® Embedded Process enables users to take advantage of SAS® In-Database Code Accelerator for Hadoop to run scoring models. By using these key SAS® technologies, the overall experience of analyzing unstructured text data can be greatly improved. The paper also provides guidelines and examples on how to deploy and run category, concept, and sentiment models for text analytics in Hadoop.

INTRODUCTION

With the rapid development of digital information technologies, enterprises are able to collect and store large amounts of data. The data size could be as large as terabytes and petabytes. Among the collected data, unstructured (primarily text) data accounts for more than 80 percent and is growing at an exponential rate. The unstructured content can be found from different sources, for example, blogs, forums, daily logs, product reviews, call center logs, emails, customer reviews and other forms of electronic text. Opportunities for analyzing such textual data to reveal insightful information for improving business operations and performance are attractive. Text analytics is the mechanism that enables organizations to automate the identification of topics, entities, facts, and events, coupled with sentiment and other subjective information.

In order to analyze the volume of data without consuming an excessive amount of network resources, IT professionals seek tools that empower them to deploy analytical algorithms inside of their Hadoop infrastructure.

In this paper, we presents an efficient and seamless approach to deploy and score SAS Text Analytics models in Hadoop. By using key SAS technologies, large amounts of unstructured text data are analyzed and harnessed so that you can understand the value of the data. SAS Contextual Analysis lets the user customize and build text analytics scoring models. These models can then be published to a Hadoop cluster. SAS DS2 is the programming language used to execute the text analytics models in Hadoop from the client machine. SAS In-Database Code Accelerator for Hadoop and SAS Embedded Process enable the DS2 program to run in Hadoop.

This paper is structured as follows:

- The basic components and framework of SAS Contextual Analysis In-Database Scoring for Hadoop technology are discussed.
- A concrete example to show how you can use the SAS DS2 program and text analytics models to score text data is provided.

OVERVIEW OF SAS® CONTEXTUAL ANALYSIS IN-DATABASE SCORING FOR HADOOP

As shown in Figure 1, SAS Contextual Analysis generates binary text analytics models and DS2 score code. The saved binary models are deployed to Hadoop cluster data nodes to analyze data in Hadoop. The extracted DS2 code requires modifications in order to satisfy the configuration of specific Hadoop distribution versions and the location of deployed text analytics models. Details of DS2 modifications are explained in Section: Running SAS Text Analytics in Hadoop.

The SAS In-Database Code Accelerator for Hadoop enables you to publish a DS2 threaded program and its associated files to the database, and then execute the program in parallel within SAS Embedded Process. When the DS2 code is executed, a MapReduce job is created and run using the YARN resource manager. The MapReduce job uses the SAS Text Analytics components in SAS Embedded Process, which runs where the data resides.

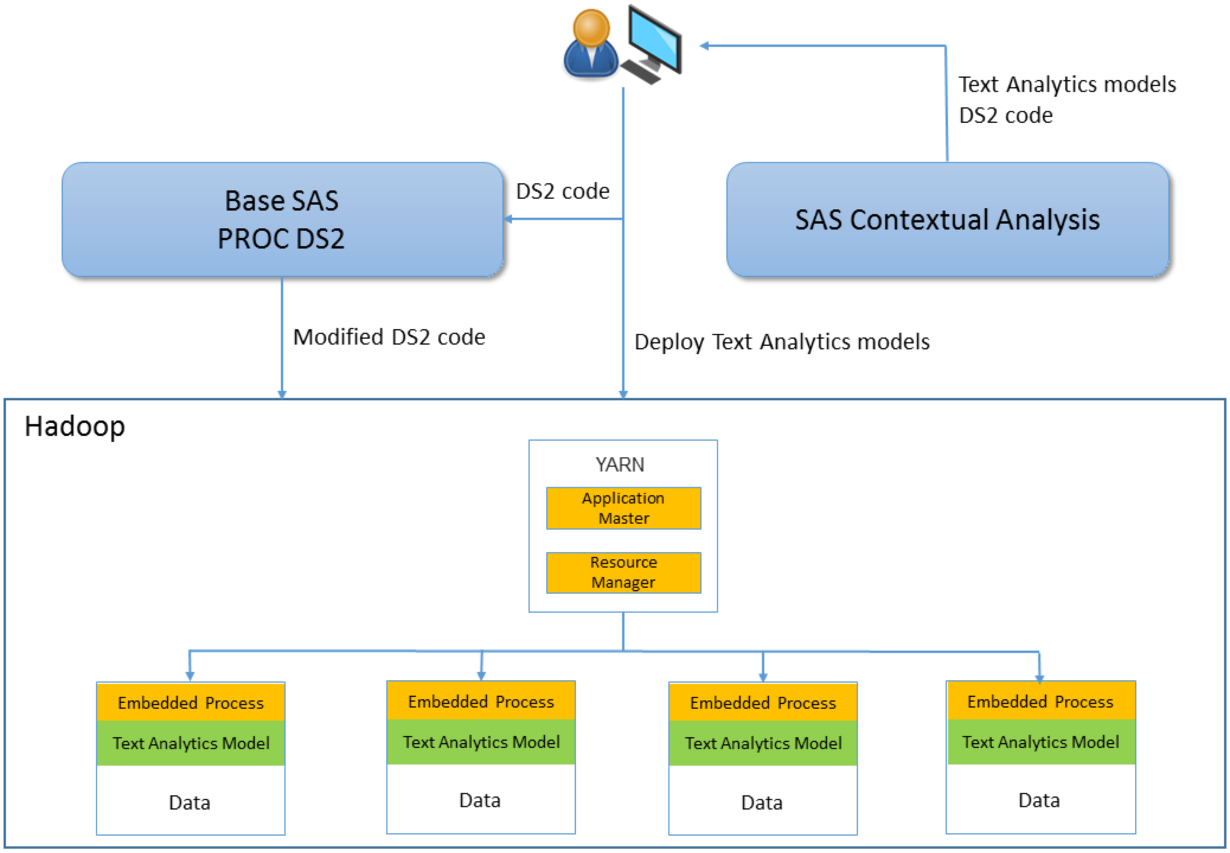


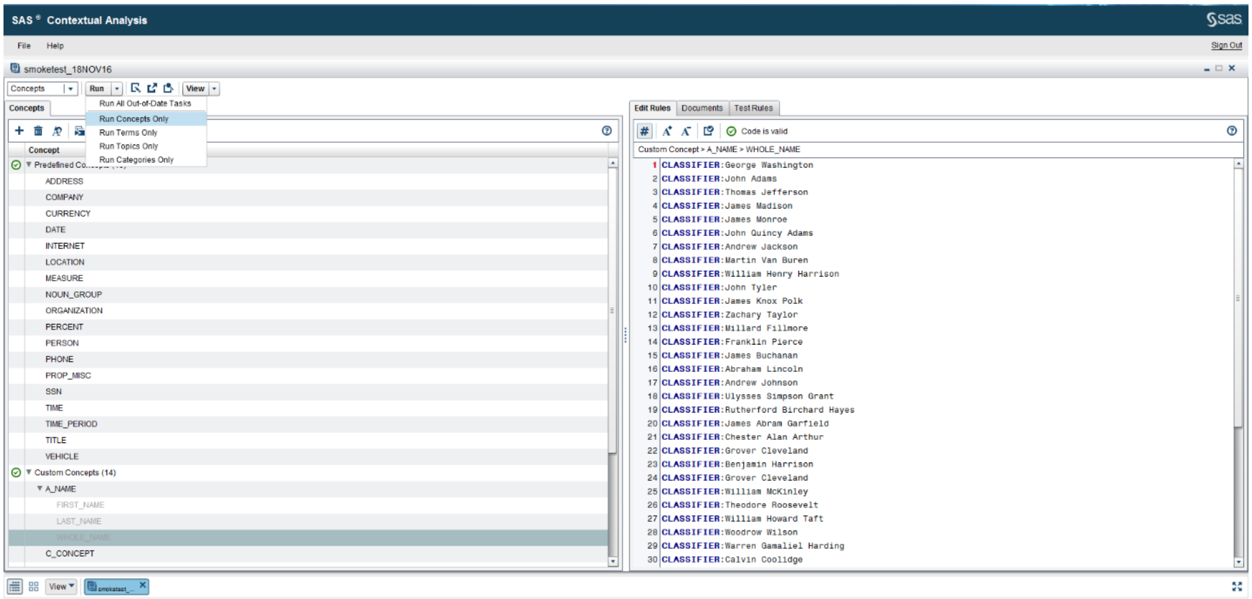
Figure 1. Framework of SAS Contextual Analysis In-Database Scoring for Hadoop

SAS CONTEXTUAL ANALYSIS

SAS Contextual Analysis is a web-based text analytics application that uses natural language processing and machine learning to derive insights from textual data. Using this application, you can determine topic identification, categorization, entity and fact extraction, and sentiment analysis in a single user interface. The application enables you to build models (based on training documents) and create taxonomies and rule sets to analyze documents. You can then customize your models for your business domain in order to realize the value of your text-based data (Bultman 2016).

At the end of the modeling process, SAS Contextual Analysis generates DS2 code and binary text analytics models for scoring text data. SAS Contextual Analysis generates three types of DS2 score code and models corresponding to categorization, concept extraction, and sentiment analysis. The DS2 code can be run within a SAS environment such as SAS® Studio or within Hadoop using SAS In-Database Code Accelerator for Hadoop. The binary models represent the rule sets for categorization (file extension: .mco), concepts (file extension: .li) and sentiment (file extension: .sam) taxonomies, which are highly optimized to apply all rules in parallel.

Display 1 shows the concept model build options in SAS Contextual Analysis.



```
CLASSIFIER:George Washington
CLASSIFIER:John Adams
CLASSIFIER:Thomas Jefferson
CLASSIFIER:James Madison
CLASSIFIER:James Monroe
CLASSIFIER:John Quincy Adams
CLASSIFIER:Andrew Jackson
CLASSIFIER:Martin Van Buren
CLASSIFIER:William Henry Harrison
CLASSIFIER:John Tyler
CLASSIFIER:James Knox Polk
CLASSIFIER:Zachary Taylor
CLASSIFIER:Millard Fillmore
CLASSIFIER:Franklin Pierce
CLASSIFIER:James Buchanan
CLASSIFIER:Abraham Lincoln
CLASSIFIER:Andrew Johnson
CLASSIFIER:Ulysses Simpson Grant
CLASSIFIER:Rutherford Birchard Hayes
CLASSIFIER:James Abram Garfield
CLASSIFIER:Chester Alan Arthur
CLASSIFIER:Grover Cleveland
CLASSIFIER:Benjamin Harrison
CLASSIFIER:Grover Cleveland
CLASSIFIER:William McKinley
CLASSIFIER:Theodore Roosevelt
CLASSIFIER:William Howard Taft
CLASSIFIER:Woodrow Wilson
CLASSIFIER:Warren Gamaliel Harding
CLASSIFIER:Calvin Coolidge
```

Display 1. Building a Concept Model in SAS Contextual Analysis

DEPLOYING SAS TEXT ANALYTICS SCORING MODELS

To deploy a SAS Text Analytics scoring model generated from SAS Contextual Analysis, specific steps must be followed, as shown below. The overall process is illustrated in Figure 2.

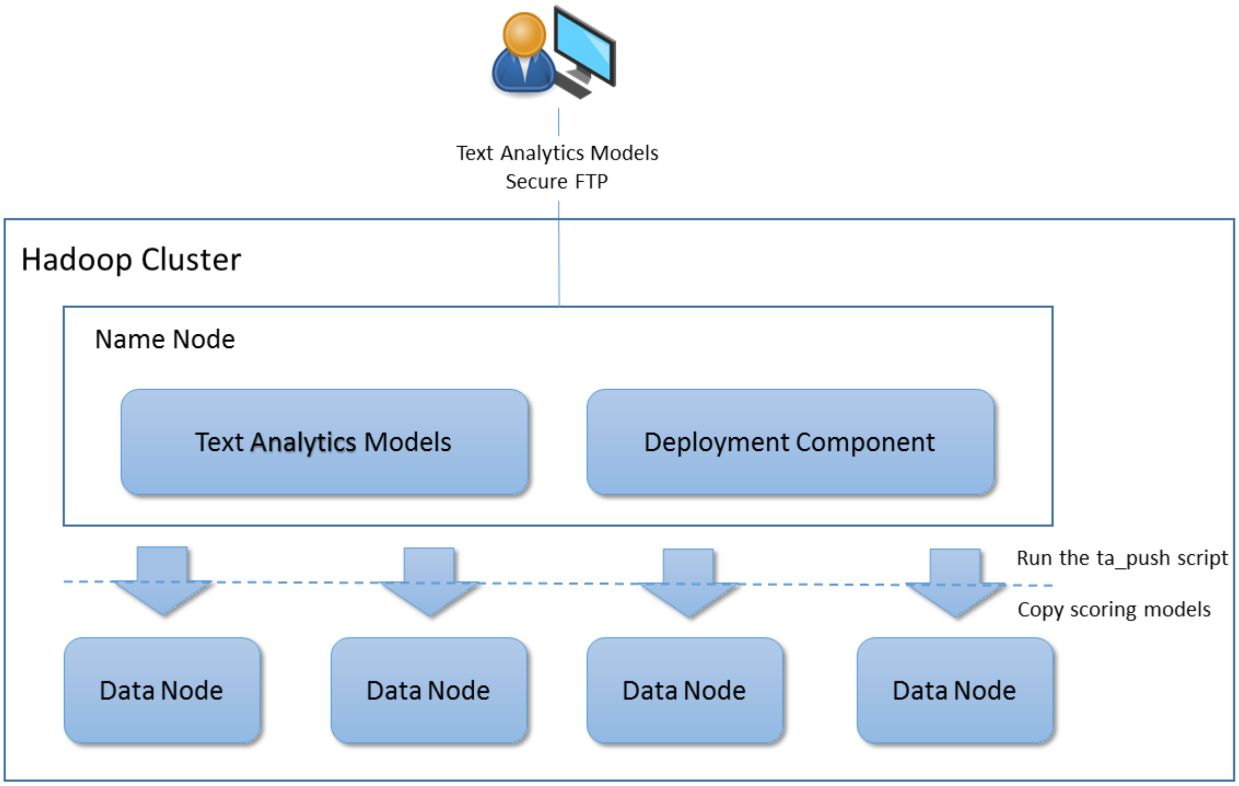


Figure 2. Deploying SAS Text Analytics Scoring Models

1. Obtain a SAS Text Analytics model for scoring concept extraction (.li file), categorization (.mco file),

or sentiment analysis (.sam file).

2. Copy the scoring code model to the Hadoop name node. After you have obtained a SAS Text Analytics model, you must copy it to the Hadoop name node. It is recommended that you copy the model to a temporary staging area, such as **/tmp/tastage**. You can copy the model to the Hadoop Name Node by using a file transfer command such as FTP or SCP.

3. Use the ta_push.sh script to deploy the SAS Text Analytics score code mode in the cluster. SAS Contextual Analysis In-Database Scoring for Hadoop provides a ta_push.sh executable file to enable you to deploy the SAS Text Analytics models on Hadoop cluster nodes. The ta_push.sh file copies the specified model to a user-specified location on each of the Hadoop nodes. The ta_push.sh file automatically discovers all nodes in the cluster and deploys the model to the specified target model path on each of the cluster data nodes.

   The ta_push.sh file must be run as the root user. The root user becomes the HDFS user in order to detect the nodes in the cluster. Here is an example of running the script:

   ```
   cd EPInstallDir/SASEPHome/bin
   ./ta_push.sh –s /tmp/tastage/en-ne.li –t /opt/sas/ta/model/en-ne.li
   ```

SAS DS2 LANGUAGE

DS2 is a procedural programming language with variables and scope, methods, packages, control flow statements, table I/O statements, and parallel programming statements. A DS2 program can be run in various SAS products, including Base SAS®, SAS® High-Performance Analytics, SAS® In-Database Code Accelerator for Hadoop, and SAS® In-Memory Analytics. The core of DS2 is similar to SAS DATA step language. Expressions in DATA step operate the same way in DS2 and most DATA step functions can be called within DS2.

DS2 has programming blocks to enable you to perform specific tasks in a modular fashion. The programming blocks are automatically invoked as DS2 executes the program.

- **Variable** - Variables are declared at the top of a DS2 program and are considered global variables having global scope (within the program).
- **Package** - The threaded-kernel (TK) extension package variables are declared at the top of the DS2 program. The package variables have global scope (within the program).
- **Methods** - DS2 classifies initialization, processing, and termination phases using three system-defined methods: init(), run(), and term(), respectively.
  - The init() method is called at the start of the program. The method is used to initialize variables and invoke the package initialization process.
  - The run() method is called after the init() method. The method iterates on all observations in the input data set.
  - The term() method is called after the run() method completes its processes. The method performs any wrap-up processing and outstanding resource cleanup.
- **Thread** - Concurrent processing is allowed using a threaded program. The threaded programming block must be declared with the `thread` command and must end with `endthread`.
- **Data program** - The block of code declared with the DATA step statement.

The programming block defines the scope of the declared variables. A variable in a nested programming block has local scope while a variable in the outermost programming block has global scope.

SAS EMBEDDED PROCESS

SAS Embedded Process is a lightweight execution container that is easily deployable on a variety of platforms. SAS Embedded Process enables DS2 code to run inside Hadoop and effectively leverages the massive parallel processing of Hadoop. The process must be installed on every node of the Hadoop cluster to capitalize on its parallelism while running DS2 code, where the process acts as a MapReduce application. Therefore, all computing resources used by SAS Embedded Process are manageable by the YARN resource manager (Ghazaleh 2016).

SAS Embedded Process includes SAS Text Analytics components where it consists of natural language processing thread-kernel (TK) extensions. The components enable you to execute text analytics scoring models in Hadoop. The same threaded-kernel extensions are installed in SAS Contextual Analysis.

RUNNING SAS TEXT ANALYTICS IN HADOOP

Understanding the DS2 code necessary to run SAS Text Analytics models inside Hadoop is simple. However, it involves understanding particular features that are related to Hadoop environment settings. This section presents the steps needed to properly configure Hadoop environment variables. It also presents a complete example of the process for writing DS2 code for scoring SAS Text Analytics.

SETTING SAS ENVIRONMENT VARIABLES

In order to run SAS Text Analytics models in Hadoop, the Hadoop Java client information must be specified in DS2 code. The Hadoop client Java Archive (JAR) files and XML client configuration files enable the client to connect the Hadoop services. Both JAR and XML files are specific to the version of the Hadoop distribution and provided by Hadoop vendors.

The Hadoop client JAR and services configuration files can be collected directly from the local file system in one of the Hadoop nodes. The configuration files are commonly stored under **/etc/hadoop/conf** (for Hadoop) and **/etc/hive/conf** (for Hive).

The Hadoop client files can also be collected using the SAS Deployment Manager. The SAS Deployment Manager requires you answer a set of questions that help locate Hadoop Java client files. For complete information about collecting Hadoop client files, see *SAS® Contextual Analysis In-Database Scoring for Hadoop: Administrator's Guide*. You must have system privileges to perform the deployment process.

After the Hadoop client JAR and services configuration files are obtained, you must set the following environment variables. The location that contains all of the Hadoop client JAR and services configuration files must be accessible from the client machine.

1. SAS_HADOOP_JAR_PATH: Specifies the location of Hadoop client JAR files. The location that contains Hadoop client JAR files must be accessible on the client machine. The variable is set by using an OPTION SET statement. The following shows sample code to set the Hadoop JAR location.

   ```
   options set=SAS_HADOOP_JAR_PATH=”C:\your\Hadoop\jars\location”;
   ```

2. SAS_HADOOP_CONFIG_PATH: Specifies the location of Hadoop services configuration files. The location that contains Hadoop services configuration files must be accessible on the client machine. The variable is set by using an OPTION SET statement. The following shows sample code to set the Hadoop services configuration location.

   ```
   options set=SAS_HADOOP_CONFIG_PATH=”C:\your\Hadoop\config\location”;
   ```

DS2 CODING PROCESS

The process of writing DS2 code for SAS Text Analytics models in Hadoop is outlined here.

1. Enable the SAS In-Database Code Accelerator to execute the DS2 threaded program inside Hadoop by using the following command:

   ```
   proc ds2 ds2accel=yes;
   ```

2. Set SAS environment variables that specify the location of the Hadoop Java client API and configuration files.

```
options set=SAS_HADOOP_JAR_PATH=”C:\your\Hadoop\jars\location”;

options set=SAS_HADOOP_CONFIG_PATH=”C:\your\Hadoop\config\location”;
```

3. Use a LIBNAME statement to gain access to a Hadoop cluster to read and write data to and from Hadoop. The LIBNAME statement is used to assign a library reference to associate with a Hadoop HDFS or Hive server. SAS in-database technologies use the SAS/ACCESS Interface to Hadoop. SAS/ACCESS Interface to Hadoop provides enterprise data access and integration between SAS and Hadoop and works similarly to other SAS engines.

```
libname gridlib hadoop server="myhivenode.com" user=myuserid
password=mypassword;
```

4. Indicate where to find the binary files that are installed during the deployment of SAS Text Analytics scoring models. Depending on what scoring models you are using, look for one of following.

```
/** Concept model scoring binary**/
%let liti_binary_path = '/path/to/liti/binary.li';

/** Category model scoring binary**/
%let mco_binary_path = '/path/to/mco/binary.mco';

/** Sentiment model scoring binary**/
%let sam_binary_path = '/path/to/sam/binary.sam';
```

5. Define a threaded program to allow parallel execution.

```
THREAD workerth / overwrite=YES;
```

CONCEPT SCORING EXAMPLE

The following DS2 code is a complete program for concept model scoring. The category and sentiment model scoring are similar. You can also find SAS Text Analytics models scoring examples in *SAS® Contextual Analysis In-Database Scoring for Hadoop: User’s Guide*.

```
/*************************************************************************
 * Set SAS environment variables that specify the location of the Hadoop
 * Java client API and configuration files. They are used to access Hadoop
 * services. The Java client API is provided by the Hadoop vendor in the
 * form of JAR files.
 *
 * NOTE: The SAS Contextual Analysis In-Database Scoring for Hadoop
 * Administrator's Guide describes how to collect Hadoop Java client API
 * JARs and configuration files.
**************************************************************************/
options set=SAS_HADOOP_JAR_PATH="C:\path\to\Hadoop\jars";
options set=SAS_HADOOP_CONFIG_PATH="C:\path\to\Hadoop\conf";

/*************************************************************************
 * Execute a LIBNAME statement to assign a library reference to associate
 * with a Hadoop HDFS or HIVE server.
 *
 * Please contact your IT administrator for the HIVE server name.
 *************************************************************************/
libname gridlib hadoop server="hivenode.com" user=userid password=password;

/*************************************************************************
 * Execute a LIBNAME statement to assign a library reference to the
 * location of the local SAS data set to be scored.
 *
 * If all the data sets used for scoring are already in HDFS, the library
 * might not be needed.
 ***************************************************************************/
libname home "C:\path\to\local\dataset";

/***************************************************************************
 * Copy the data set to HDFS. Unique observation IDs are required and are
 * created as the field _document_id.
 *
 * If all the data sets used for scoring are already in HDFS, this step
 * might not be needed.
 ***************************************************************************/
data gridlib.input_dataset;
  set home.input_dataset;
  _document_id = _N_;
Run

/***************************************************************************
 * Set input/output macro variables.
 *
 * input_ds: Name of the input data set in Hadoop, including the HDFS
libref
 * document_id: Name of the unique document ID column in the input data set
 * document_column: Name of the column to process in the input data set
 * liti_binary_path: Path to the LITI binary
 * output_ds: Name of the output data set, including the HDFS libref
 ***************************************************************************/
%let input_ds = gridlib.input_dataset;
%let document_id = _document_id;
%let document_column = text_to_process;
%let liti_binary_path = '/path/to/liti/binary.li';
%let output_ds = gridlib.concept_out;

/***************************************************************************
 * Delete the output before starting (Optional)
 ***************************************************************************/
proc delete data=&output_ds; run;

/***************************************************************************
 * Scores concepts in Hadoop
 ***************************************************************************/
proc ds2 ds2accel=yes xcode=warning;

  /* These packages are part of the Text Analytics add-on and are      */
  /* installed in the EP                                               */
  require package tkcat; run;
  require package tktxtanio; run;

  /* The output of the thread program is the input of the data program */
  THREAD workerth / overwrite=YES;

    dcl package tkcat cat();
    dcl package tktxtanio txtanio();
    dcl binary(8) _apply_settings;
    dcl binary(8) _document;
    dcl binary(8) _liti_binary;
    dcl binary(8) _trans;
    dcl double _status;
    dcl double _num_matches;
dcl double _i;
dcl double _document_id;
dcl varchar(1024) _name;
dcl varchar(1024) _full_path;
dcl double _start_offset;
dcl double _end_offset;
dcl varchar(1024) _term;
dcl varchar(1024) _canonical_form;
retain _apply_settings;
retain _liti_binary;
retain _trans;

/*********************************************************************
 * Initialization step. Only runs once when starting.
 *********************************************************************/
method init();
  _apply_settings = cat.new_apply_settings();
  _liti_binary = txtanio.new_on_content_server(&liti_binary_path);

  _status = cat.set_apply_model(_apply_settings, _liti_binary);
  if _status NE 0 then put 'ERROR: set_apply_model fails';

  /* Match types are 0=ALL, 1=LONGEST or 2=BEST */
  _status = cat.set_match_type(_apply_settings, 0);
  if _status NE 0 then put 'ERROR: set_match_type fails';

  _status = cat.initialize_concepts(_apply_settings);
  if _status NE 0 then put 'ERROR: initialize_concepts fails';

  _trans = cat.new_transaction();
end;

/*********************************************************************
 * Run step. The method runs per row of input.
 *********************************************************************/
method run();
  set &input_ds(keep=(&document_column &document_id));

  /* Only process if document observation is not empty*/
  if &document_column NE ' ' then do;

    /* Initialize the document with the column data */
    _document = txtanio.new_document_from_string(&document_column);

    /* Set the document on the transaction so we're ready to process */
    _status = cat.set_document(_trans, _document);
    if _status NE 0 then put
      'ERROR: set_document fails on obs:' &document_id;

    /* Apply the binary to the document */
    _status = cat.apply_concepts(_apply_settings, _trans);
    if _status NE 0 then put
      'ERROR: apply_concepts fails on obs:' &document_id;

    /* Look for the concept matches */
    _num_matches = cat.get_number_of_concepts(_trans);
    _i = 0;
```

```
        do while (_i LT _num_matches);
          _name = cat.get_concept_name(_trans, _i);
          _full_path = cat.get_full_path_from_name(_trans, _name);
          _start_offset = cat.get_concept_start_offset(_trans, _i);
          _end_offset = cat.get_concept_end_offset(_trans, _i);
          _term = cat.get_concept(_trans, _i);
          _canonical_form = cat.get_concept_canonical_form(_trans, _i);

          output;

          _i = _i + 1;
        end;

        /* Now look for fact matches */
        _num_matches = cat.get_number_of_facts(_trans);
        _i = 0;
        _canonical_form = '';
        do while (_i LT _num_matches);
          _name = cat.get_fact_name(_trans, _i);
          _full_path = cat.get_full_path_from_name(_trans, _name);
          _start_offset = cat.get_fact_start_offset(_trans, _i);
          _end_offset = cat.get_fact_end_offset(_trans, _i);
          _term = cat.get_fact(_trans, _i);

          output;

          _i = _i + 1;
        end;
        _i = 0;

        /* Clean up resources */
        cat.clean_transaction(_trans);
        txtanio.free_object(_document);
      end;
    end;

    /********************************************************************
     * Termination step that runs only once at the end.
     ********************************************************************/
    method term();
      /* clean up resources */
      cat.free_transaction(_trans);
      cat.free_apply_settings(_apply_settings);
      txtanio.free_object(_liti_binary);
    end;
  endthread;
run;

/************************************************************************
 * Collect output data
 ************************************************************************/
data &output_ds(
  keep=(
    &document_id
    _name
    _full_path
    _start_offset
    _end_offset
    _term
    _canonical_form
  )
  overwrite=yes
);
dcl THREAD workerth THRD;

method run();
  set from THRD;
end;

enddata;
run; quit;
```

CONCLUSION

SAS Contextual Analysis In-Database Scoring for Hadoop technology provides the SAS user with a powerful framework to deploy and run category, concept, and sentiment scoring models for text analytics in Hadoop. The framework enables you to seamlessly deploy SAS Text Analytics scoring models into a Hadoop cluster. The deployed models are then used to score unstructured text data using SAS DS2. In this paper, we discussed the overall process of SAS Contextual Analysis In-Database Scoring for Hadoop and its basic components. We also demonstrated steps of how a SAS user can modify the DS2 code to score in Hadoop.

REFERENCES

Ghazaleh, David. 2016. “Exploring SAS® Embedded Process Technologies on Hadoop.” *Proceedings of the SAS Global 2016 Conference*, Las Vegas, NV: Available at http://support.sas.com/resources/papers/proceedings16/SAS5060-2016.pdf.

Bultman, David. 2016. “Running Text Analytics Models in Hadoop.” *Proceedings of the SAS Global 2016 Conference*, Las Vegas, NV: Available at http://support.sas.com/resources/papers/proceedings16/SAS4880-2016.pdf.

ACKNOWLEDGMENTS

David Bultman, Principal Software Developer, Cloud Analytic Services, SAS Institute Inc.

Gail Pearson, Sr. Technical Writer, SAS Institute Inc.

Laura Rock, Manager, Advanced Analytics Testing, SAS Institute Inc.

Peggy James, Sr. Analytics Software Tester, SAS Institute Inc.

RECOMMENDED READING

- *Paper SAS4880-2016, “Running Text Analytics Models in Hadoop” by David Bultman and Adam Pilz, SAS Institute, Inc.*
- *Paper SAS5060-2016, “Exploring SAS® Embedded Process Technologies on Hadoop” by David Ghazaleh, SAS Institute, Inc.*
- *SAS® Contextual Analysis: User’s Guide*
- *SAS® Contextual Analysis In-Database Scoring for Hadoop: Administrator’s Guide*
- *SAS® Contextual Analysis In-Database Scoring for Hadoop: User’s Guide*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Seung Lee
SAS Institute, Inc.
(919) 531-6659
SeungYong.Lee@sas.com

Xu Yang
SAS Institute, Inc.
(919) 531-3421
Xu.Yang@sas.com

Saratendu Sethi
SAS Institute, Inc.
(919) 531-0597
Saratendu.Sethi@sas.com

Paper SAS282-2017

Applying Text Analytics and Machine Learning to Assess Consumer Financial Complaints

Tom Sabo, SAS Institute Inc., Arlington, VA

ABSTRACT

The Consumer Financial Protection Bureau (CFPB) collects tens of thousands of complaints against companies each year, many of which result in the companies in question taking action, including making payouts to the individuals who filed the complaints. Given the volume of the complaints, how can an overseeing organization quantitatively assess the data for various trends, including the areas of greatest concern for consumers?

In this paper, we apply a repeatable model of text analytics techniques to the publicly available CFPB data. Specifically, we use SAS® Contextual Analysis to explore sentiment and machine learning techniques to model the natural language available in each free-form complaint against a disposition code for the complaint, primarily focusing on whether a company paid out money. This process generates a taxonomy in an automated manner. We also explore methods to structure and visualize the results, showcasing how areas of concern are made available to analysts using SAS® Visual Analytics and SAS® Visual Statistics. Finally, we discuss the applications of this methodology for overseeing government agencies and financial institutions alike.

INTRODUCTION

The CFPB is one of a number of overseeing institutions that ensure that the consumer is treated fairly by corporations and financial institutions. It stands alongside other international organizations such as the United Kingdom Financial Conduct Authority, the Australian Competition and Consumer Commission, and EEC-Net, which assists European Union consumers in resolving international purchase complaints. The CFPB was established in 2011. It was created after the financial crisis of 2008 to help consumers resolve problems at the transactional level, and to address larger macro-level issues before they become unmanageable. The CFPB is responsible for more than 11.7 billion dollars of relief for consumers due to enforcement actions[1].

The CFPB has handled more than one million complaints since its inception, and this number is increasing annually[1]. As more complaints are filed, is the solution to handling the increasing workload adding more readers to manually address the complaints and identify trends? Generally speaking, scaling up manual analysis of textual data has three challenges. First, unless very specific standards (bordering on definitive rules) are adopted, the method that one reader uses to address and tag a complaint can be quite different from the method a second reader uses. Scale this difference up to many readers, and you have many different, qualitative interpretations of the textual data. Second, reader fatigue ensures that the way a reader will address the first 10 complaints of the day will not necessarily be the same as the way they address the last 10 complaints. Vital information might be missed or skipped. Finally, suppose a trend is uncovered, and the directive arises to go back and retag all the data from the past year with this new trend. This is a case where manual analysis doesn’t scale, and often enough, a simple search operation for a trend pattern will not be sufficient.

The benefit for potential analysis of the CFPB data is that each of the records is tagged with a disposition code, denoting the action taken by the organization against which the complaint was filed. With this information (and to a lesser extent, independent of it), we can uncover trends surrounding the actions taken (for example, what were the defining characteristics of complaints where the organization in question paid out monetary compensation to the individuals filing the complaints vs. complaints with a lower disposition such as those closed simply with an explanation?) In the sections that follow in this paper, we will explore a short end-to-end implementation that showcases how an analyst can use SAS technology to quantitatively assess the complaint data for various trends. This includes the consumers’ areas of greatest concern, as well as areas of complaint that are in need of legislative correction. We show how to apply a sentiment model to the text as well as machine learning methods through SAS

Contextual Analytics to accomplish this. Finally, we will assess the results using visualization capabilities to highlight actionable information.

Specifically, we will apply a process built upon three previously presented papers at SAS Global Forum: one in 2014 to define a framework for research analytics[2], a second in 2015 to extend this framework for government spending[3], and a third in 2016 to apply the framework to auto-categorization of event data in conflict affected regions[4]. We encourage the reader to refer back to these papers to gain a sense for the wide applicability of these capabilities across several public sector use cases.

The five-step process for generating and using the framework is as follows:

1. **Data acquisition and preparation for text analytics**: Data is acquired for our example use case through web interfaces and is converted into a SAS data set using SAS® Enterprise Guide®.
2. **Text analytics**: We use SAS Contextual Analysis for sentiment analysis, as well as for modeling and rule-building techniques to generate hierarchical categorical data. This newly generated sentiment and categorical data serves as additional structured information for subsequent analysis and visualization against the CFPB data set.
3. **Data preparation for visual analysis**: We first use categorical scoring code outlined in the 2016 paper mentioned above[4]. To add a layer of sentiment information, we leverage the code provided by SAS Contextual Analysis and merge with the categorically scored table. This enables hierarchical exploration of the data in the subsequent visualization steps.
4. **Ad hoc exploration and modeling**: This is accomplished with SAS Visual Analytics and SAS Visual Statistics.
5. **Interactive report generation and use**: This is also accomplished with SAS Visual Analytics.

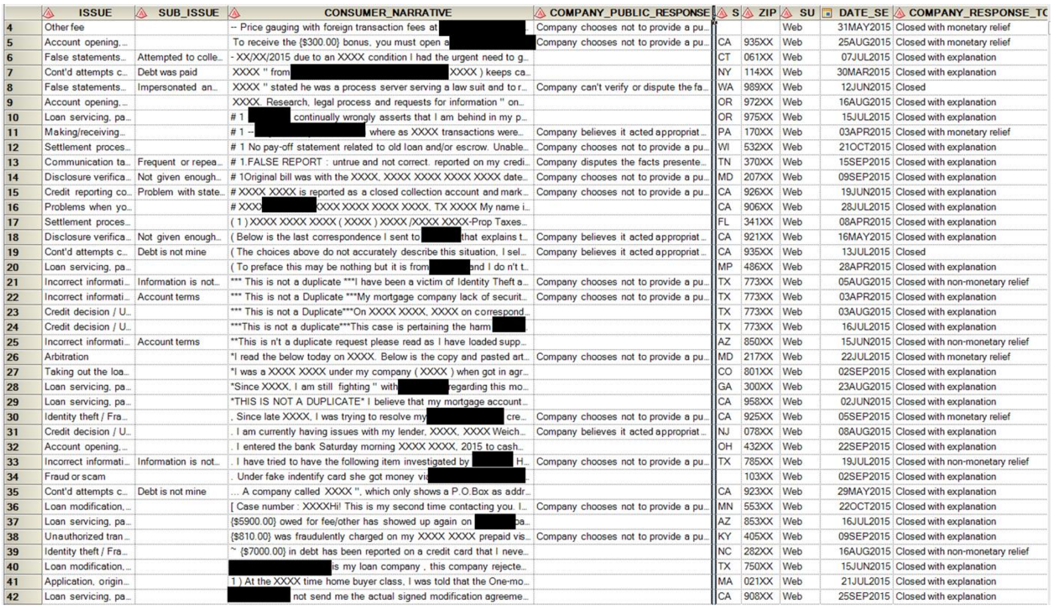
DATA ACQUISITION AND PREPARATION FOR TEXT ANALYTICS

We obtained the CFPB data from the interface available on its publicly facing website[5] and specified that only data with a narrative should be pulled. For the project, we used data from March 19, 2015 to October 30, 2015. This amounted to 37,619 complaints with a narrative. We imported the data into a SAS data set using SAS Enterprise Guide. In the process, we retained the original SAS data set of 37,619 complaints, but we also generated a new representative data set of 15K complaints for the interactive and modeling-based text analytics work.

TEXT ANALYTICS – SAS CONTEXTUAL ANALYTICS

After registering the SAS data set in metadata using SAS® Management Console, we select and load the data set within the SAS Contextual Analysis interface. For this project we selected the option within the interface to run a document-level (complaint level) sentiment model. In our case, we chose to run the default sentiment model. However, we could also use a specialized sentiment model developed using SAS® Sentiment Analysis Studio® in its place.

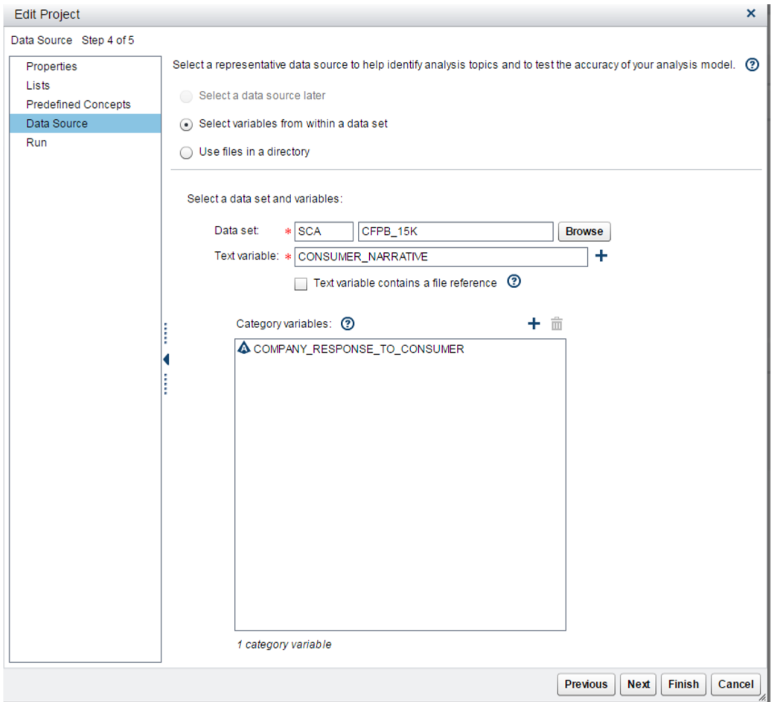
When selecting the data set, we specify the CONSUMER_NARRATIVE column as the freeform text field to perform text analytics against, as well as specifying the COMPANY_RESPONSE_TO_CONSUMER as a categorical target variable. See Figure 1 below for an example snippet of the data, including additional structured data columns. The names of the various financial and retail organizations called out in the narrative complaints have been obscured. Also note that the CFPB has also already obfuscated all personally identifiable information for each consumer using XXXX notation.



	ISSUE	SUB_ISSUE	CONSUMER_NARRATIVE	COMPANY_PUBLIC_RESPONSE	S	ZIP	SU	DATE_SE	COMPANY_RESPONSE_TC
4	Other fee		-- Price gauging with foreign transaction fees at [redacted].	Company chooses not to provide a pu...			Web	31MAY2015	Closed with monetary relief
5	Account opening, ...		To receive the {$300.00} bonus, you must open a [redacted]	Company chooses not to provide a pu...	CA	935XX	Web	25AUG2015	Closed with monetary relief
6	False statements...	Attempted to colle...	- XX/XX/2015 due to an XXXX condition I had the urgent need to g...		CT	061XX	Web	07JUL2015	Closed with explanation
7	Cont'd attempts c...	Debt was paid	XXXX '' from [redacted] XXXX) keeps ca...		NY	114XX	Web	30MAR2015	Closed with explanation
8	False statements...	Impersonated an...	XXXX '' stated he was a process server serving a law suit and to r...	Company can't verify or dispute the fa...	WA	989XX	Web	12JUN2015	Closed
9	Account opening, ...		XXXX. Research, legal process and requests for information '' on...		OR	972XX	Web	16AUG2015	Closed with explanation
10	Loan servicing, pa...		# 1 [redacted] continually wrongly asserts that I am behind in my p...		OR	975XX	Web	15JUL2015	Closed with explanation
11	Making/receiving...		# 1 -- [redacted] where as XXXX transactions were...	Company believes it acted appropriat ...	PA	170XX	Web	03APR2015	Closed with monetary relief
12	Settlement proces...		# 1 No pay-off statement related to old loan and/or escrow. Unable...	Company chooses not to provide a pu...	WI	532XX	Web	21OCT2015	Closed with explanation
13	Communication ta...	Frequent or repea...	# 1.FALSE REPORT : untrue and not correct. reported on my credi...	Company disputes the facts presente...	TN	370XX	Web	15SEP2015	Closed with explanation
14	Disclosure verifica...	Not given enough...	# 1Original bill was with the XXXX, XXXX XXXX XXXX XXXX date...	Company chooses not to provide a pu...	MD	207XX	Web	09SEP2015	Closed with explanation
15	Credit reporting co...	Problem with state...	# XXXX XXXX is reported as a closed collection account and mark...	Company chooses not to provide a pu...	CA	926XX	Web	19JUN2015	Closed with explanation
16	Problems when yo...		# XXX[redacted]XXX XXXX XXXX XXXX, TX XXXX My name i...		CA	906XX	Web	28JUL2015	Closed with explanation
17	Settlement proces...		(1) XXXX XXXX XXXX (XXXX) XXXX /XXXX XXXX-Prop Taxes...		FL	341XX	Web	08APR2015	Closed with explanation
18	Disclosure verifica...	Not given enough...	(Below is the last correspondence I sent to [redacted] that explains t...	Company believes it acted appropriat ...	CA	921XX	Web	16MAY2015	Closed with explanation
19	Cont'd attempts c...	Debt is not mine	(The choices above do not accurately describe this situation, I sel...	Company believes it acted appropriat ...	CA	935XX	Web	13JUL2015	Closed
20	Loan servicing, pa...		(To preface this may be nothing but it is from [redacted] and I do n't t...		MP	486XX	Web	28APR2015	Closed with explanation
21	Incorrect informati...	Information is not...	*** This is not a duplicate ***I have been a victim of Identity Theft a...	Company chooses not to provide a pu...	TX	773XX	Web	05AUG2015	Closed with non-monetary relief
22	Incorrect informati...	Account terms	*** This is not a Duplicate ***My mortgage company lack of securit...	Company chooses not to provide a pu...	TX	773XX	Web	03APR2015	Closed with explanation
23	Credit decision / U...		*** This is not a Duplicate***On XXXX XXXX, XXXX on correspond...		TX	773XX	Web	03AUG2015	Closed with explanation
24	Credit decision / U...		***This is not a duplicate***This case is pertaining the harm [redacted].		TX	773XX	Web	16JUL2015	Closed with explanation
25	Incorrect informati...	Account terms	**This is n't a duplicate request please read as I have loaded supp...		AZ	850XX	Web	15JUN2015	Closed with non-monetary relief
26	Arbitration		*I read the below today on XXXX. Below is the copy and pasted art...	Company chooses not to provide a pu...	MD	217XX	Web	22JUL2015	Closed with monetary relief
27	Taking out the loa...		*I was a XXXX XXXX under my company (XXXX) when got in agr...		CO	801XX	Web	02SEP2015	Closed with explanation
28	Loan servicing, pa...		*Since XXXX, I am still fighting '' with [redacted] regarding this mo...		GA	300XX	Web	23AUG2015	Closed with explanation
29	Loan servicing, pa...		*THIS IS NOT A DUPLICATE* I believe that my mortgage account...		CA	958XX	Web	02JUN2015	Closed with explanation
30	Identity theft / Fra...		. Since late XXXX, I was trying to resolve my [redacted] cre...	Company chooses not to provide a pu...	CA	925XX	Web	05SEP2015	Closed with monetary relief
31	Credit decision / U...		. I am currently having issues with my lender, XXXX, XXXX Weich...	Company believes it acted appropriat ...	NJ	078XX	Web	08AUG2015	Closed with explanation
32	Account opening, ...		. I entered the bank Saturday morning XXXX XXXX, 2015 to cash...		OH	432XX	Web	22SEP2015	Closed with explanation
33	Incorrect informati...	Information is not...	. I have tried to have the following item investigated by [redacted] H...	Company chooses not to provide a pu...	TX	785XX	Web	19JUL2015	Closed with non-monetary relief
34	Fraud or scam		. Under fake indentify card she got money vi[redacted].			103XX	Web	02SEP2015	Closed with explanation
35	Cont'd attempts c...	Debt is not mine	... A company called XXXX '', which only shows a P.O.Box as addr...		CA	923XX	Web	29MAY2015	Closed with explanation
36	Loan modification, ...		[Case number : XXXXHi! This is my second time contacting you. I...	Company chooses not to provide a pu...	MN	553XX	Web	22OCT2015	Closed with explanation
37	Loan servicing, pa...		{$5900.00} owed for fee/other has showed up again on [redacted]pa...		AZ	853XX	Web	16JUL2015	Closed with explanation
38	Unauthorized tran ...		{$810.00} was fraudulently charged on my XXXX XXXX prepaid vis...	Company chooses not to provide a pu...	KY	405XX	Web	09SEP2015	Closed with explanation
39	Identity theft / Fra...		~ {$7000.00} in debt has been reported on a credit card that I neve...		NC	282XX	Web	16AUG2015	Closed with non-monetary relief
40	Loan modification, ...		[redacted] is my loan company , this company rejecte...		TX	750XX	Web	15JUN2015	Closed with explanation
41	Application, origin...		1) At the XXXX time home buyer class, I was told that the One-mo...		MA	021XX	Web	21JUL2015	Closed with explanation
42	Loan servicing, pa...		[redacted] not send me the actual signed modification agreeme...		CA	908XX	Web	25SEP2015	Closed with explanation

Figure 1: Sampling of Complaint Data Including Consumer Narrative and Company Response to Consumer

We specify a target category variable of COMPANY_RESPONSE_TO_CONSUMER as part of the SAS Contextual Analysis project definition to tell SAS to model Boolean textual rules against the CONSUMER_NARRATIVE. These textual rules differentiate term and phrase combinations that appear in each category value from other term and phrase combinations that appear in the other category values. In the context of this data set, one of the category values for COMPANY_RESPONSE_TO_CONSUMER is “Closed with Monetary Relief”. SAS Contextual Analysis can tell us what common terms, phrases, and term-phrase combinations are most often associated with monetary relief, but not typically not associated with the other category values, such as “Closed with Explanation”. Two examples of these phrases or terms are mentions of specific retail organizations not mentioned anywhere in the structured data, or the term GFE (Good Faith Estimates). This automated rule-building technology helps the analyst by characterizing the complaints that result in monetary relief, and backs up the analyst with quantitative analysis. A researcher could manually generate a similar categorical taxonomy to capture these instances, but creating this taxonomy from scratch is highly time consuming, compared to the method just described, which produces results in minutes. In addition, the researcher is likely unaware of all the patterns in the textual data. The power of the approach presented here is that it automatically generates a taxonomy that fits and describes each data set, and that taxonomy can subsequently be refined using subject matter expertise. Refining a taxonomy makes much better use of the subject matter expert’s time and resources than creating a taxonomy from scratch. Figure 2 illustrates the process of generating a project against the data using the COMPANY_REPONSE_TO_CONSUMER as a categorical variable and the CONSUMER_NARRATIVE as the text variable.



Data Source Step 4 of 5

Select a representative data source to help identify analysis topics and to test the accuracy of your analysis model.

○ Select a data source later

● Select variables from within a data set

○ Use files in a directory

Select a data set and variables:

Data set: * SCA CFPB_15K

Text variable: * CONSUMER_NARRATIVE

□ Text variable contains a file reference

Category variables:

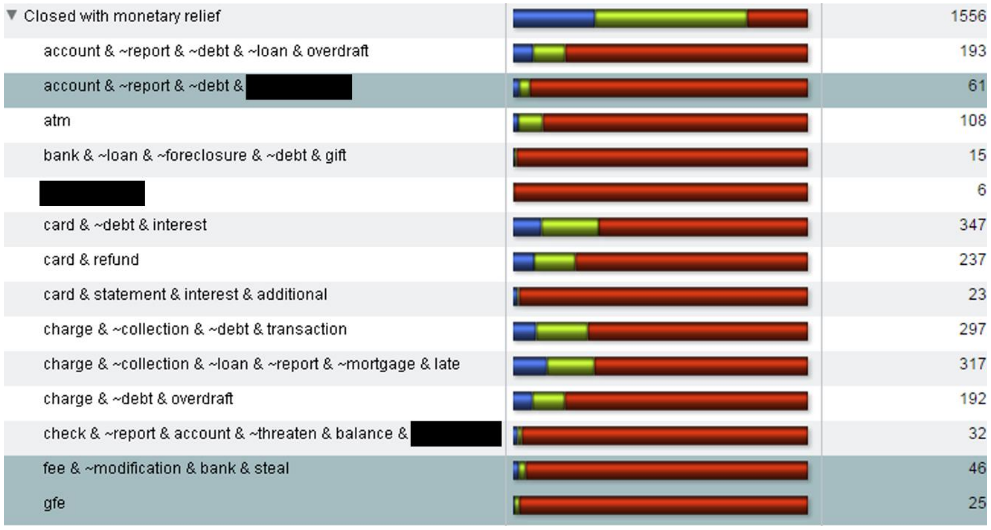
COMPANY_RESPONSE_TO_CONSUMER

1 category variable

Figure 2: Defining a New Project in SAS Contextual Analytics, Including a Category Variable

SAS Contextual Analysis includes a number of exploratory capabilities, including term and topic exploration. In this paper, we focus on its capability to generate textual rules against categorical data, and the subsequent scoring and augmentation of the original data set using these rules overlaid by a sentiment model. For a further study of terms and topic exploration in the context of a research-oriented data set, please see SAS Global Forum paper 061-2014, "Uncovering Trends in Research using Text Analytics with Examples from Nanotechnology and Aerospace Engineering." [2]

Figure 3 illustrates the textual rules that SAS Contextual Analysis auto-generates against the CONSUMER_NARRATIVE textual column for the COMPANY_RESPONSE_TO_CONSUMER value of "Closed with Monetary Relief". These rules are combinations of certain terms and phrases that appear in the narrative for complaints that result in monetary relief but that don't tend to occur for the other complaint disposition codes. These terms and phrases can be used to auto-classify new narratives as being ones that are likely to result in monetary relief, which enables analysts to prioritize the complaints that they receive. The terms and phrases are also used in subsequent steps to characterize and sub-divide the various terminology that surrounds financial compensation, which enables the analyst to explore these divisions separately and identify trends and patterns in monetary relief.



▼ Closed with monetary relief	1556
account & ~report & ~debt & ~loan & overdraft	193
account & ~report & ~debt & [redacted]	61
atm	108
bank & ~loan & ~foreclosure & ~debt & gift	15
[redacted]	6
card & ~debt & interest	347
card & refund	237
card & statement & interest & additional	23
charge & ~collection & ~debt & transaction	297
charge & ~collection & ~loan & ~report & ~mortgage & late	317
charge & ~debt & overdraft	192
check & ~report & account & ~threaten & balance & [redacted]	32
fee & ~modification & bank & steal	46
gfe	25

Figure 3: Rules Generated by SAS Contextual Analysis Related to Monetary Relief

Each Boolean rule consists of terms and phrases joined by “&,” indicating “AND,” as well as modified by a “~,” indicating “NOT.” In addition, each of the terms presented include all stemmed versions of those terms. For example, the term “steal” is representative of “stole” and “stolen” as well. Putting it all together in an example, the rule “fee & ~modification & bank & steal” indicates a complaint containing all of the terms “fee”, “bank”, and “steal”, or different stemmed versions of these, so “fees”, “banks”, and “stolen” would suffice. Also, the term “modification” or any of its stemmed versions must not be present, indicating that this rule primarily applies outside of the mortgage process, where loan modifications are common in the complaint data.

The Boolean rules are represented by a colored bar, which includes blue, yellow, and red components. The blue component of the bar represents cases where a rule correctly matches the given event type. These are true positives. The yellow component of the bar represents cases where the rule also matches for a different event type. These are false positives. The red component of the bar is primarily applicable to the “Closed with Monetary Relief” level, rather than the individual rule level. At the “Closed with Monetary Relief” level, the red component of the bars represents cases where SAS Contextual Analysis is unable to define a consistent rule to differentiate these complaint dispositions from the other disposition types such as “Closed with Explanation” or “Closed with Non-Monetary Relief”.

Rules that are generated against the narrative take a variety of forms when interpreted, all of which are indicative of some trend related to monetary compensation following a consumer complaint. They can be indicative of retail organizations that are supported by financial organizations via a company debit card, for example. This can be powerful information, because no retail organization is named anywhere in the structured complaint data. However, if the customer service for the card in question is very poor for say, ACME retail organization, individuals are apt to complain about ACME in their free-form complaints, and this connection can be made only through text analysis of the free-form narrative.

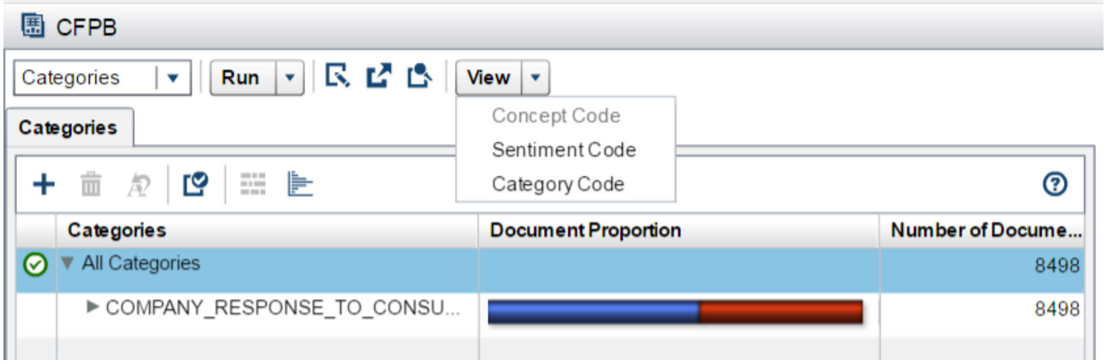
Rules are also indicative of problems with the supporting financial organizations themselves, such as account trouble that is related to a particular bank that supports a variety of retail debit cards. This might be due to widespread poor customer service involving this bank. Finally, these rules can be indicative of bank or lending practices, such good faith estimates, which often result in some type of monetary compensation when they are mentioned in a complaint. It is possible that lending organizations are taking advantage of the complication of good faith estimate statements. They could be giving confusing or inaccurate statements in order to hide fees. This last example in particular is valuable. Text Analytics quantitatively depicts that there is a practice by lending institutions that is likely being abused or misused, and provides the opportunity for an overseeing organization to take action.

DATA PREPARATION FOR VISUAL ANALYSIS

As mentioned, we are interested in both the categorical scoring and sentiment scoring of the data set. This involves a few steps, and two segments of SAS code.

1. First, we use a SAS code segment to invoke the categorical taxonomy model generated with the COMPANY_RESPONSE_TO_CONSUMER structured data field. We score against the full data set of CFPB complaints, as opposed to the 15K sample that we used to generate the model. This scores each document for, among other things, rules indicative of complaints that lead to monetary relief. To generate this code and subsequently score our complaints data set, we used methods similar to those described in the DATA PREPARATION FOR VISUAL ANALYSIS section of the SAS Global Forum paper “Extending the Armed Conflict Location and Event Data Project with SAS® Text Analytics”[4]. This method extends the out-of-the-box score code to produce a categorical hierarchy suitable for visual exploration. Please refer to the paper for information on augmenting the out-of-the-box code provided by SAS Contextual Analysis for this purpose.

2. Second, we obtain a code snippet in SAS Contextual Analysis by selecting the Sentiment Code option from the View drop-down menu. As discussed previously, this option uses a generalized sentiment model, but it could also leverage a model built from SAS® Sentiment Analysis Studio. The out-of-the-box DS2 code needs to be modified only slightly from its original format to designate input and output SAS data sets. Figure 4 shows how to access the sentiment scoring code in SAS Contextual Analysis, which is subsequently modified in a SAS programming environment.



Categories	Document Proportion	Number of Docume...
All Categories		8498
COMPANY_RESPONSE_TO_CONSU...		8498

Figure 4: Option to Depict Sentiment Code in SAS Contextual Analysis

In order to define input file locations for the environment used in this project, we modify the early lines of the out-of-the-box sentiment code to look like the following:

```
/******************************************************************
* SAS Contextual Analysis
* Sentiment Score Code
*
* Modify the following macro variables to match your needs.
*****************************************************************/
/* check if the variables were defined elsewhere - this is used for
 embedding code into SAS Text Miner */
%sysfunc(ifc(%symexist(tm_defined_vars),, %nrstr(
/* the path to the directory containing the data set you would like to
 score */
%let lib_path= D:\data\sca;
/* the data set you would like to score */
%let input_ds = _my_lib.cfpb_full;
/* the column in the data set that contains the text data to score */
%let document_column = CONSUMER_NARRATIVE;
)));
```

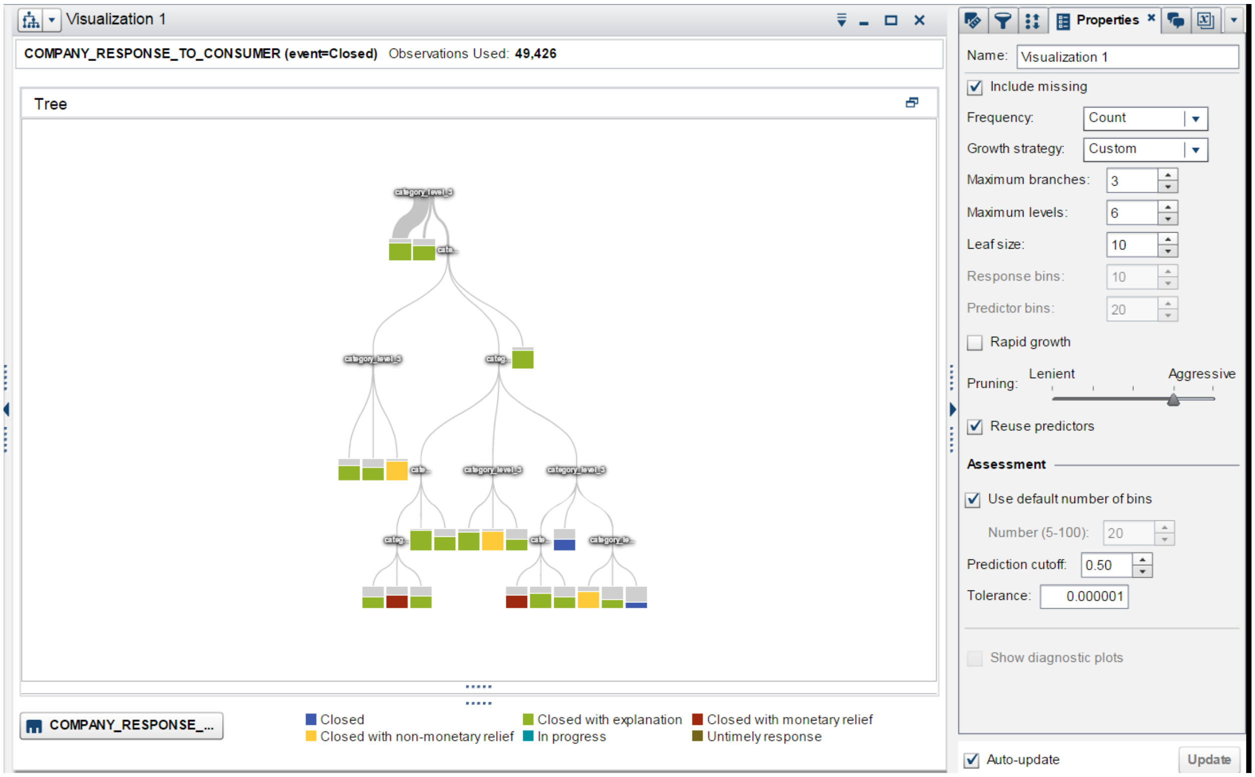
In order to save the output sentiment data for the environment used in this project, we add the following lines to the end of the sentiment code:

```
libname outputlib 'D:\data\sca\out';
data outputlib.cfpb_sentiment_scored; set &output_ds;
run;
```

3. Finally, using SAS Enterprise Guide, we simply join all the fields of the categorical results table from step 1 above with the _sentiment_probability_ field from step 2. We join for every row from table 1 against the ID column, which is present in both tables.

AD HOC EXPLORATION AND MODELING

We load the SAS data set of complaints, which includes newly generated hierarchical category scoring and document level sentiment, into SAS Visual Analytics for exploration, modeling, and reporting. SAS Visual Statistics, a set of predictive capabilities within SAS Visual Analytics, provides interactive decision trees that illuminate differentiating trends in the data. In this example, we use a decision tree to highlight textual rule combinations that are indicative of various disposition codes, such as monetary relief. To do this, we set the COMPANY_RESPONSE_TO_CONSUMER as a target, and use only the textual rules generated from the text analytics exercise associated with each event as input to the model. This will highlight branches of the decision tree where various phrases present in the narrative correlate with the different disposition codes of COMPANY_RESPONSE_TO_CONSUMER. If you re-create this example, you should also consider using the pre-existing structured data that is associated with each complaint in conjunction with the newly generated structured rules. These combinations also yield illuminating results. See Figure 5 for a high-level depiction of the generated tree, whose resulting bins characterize the data in meaningful ways. In particular, it is interesting to note branches of the tree that result in predominant dispositions other than “Closed with Explanation”, which is the overall predominant disposition. Figure 6 zooms in on the top of the decision tree to visually depict how, in general, complaints that result in monetary relief, denoted by a red bar, are significantly less frequent than ones with a “Closed with Explanation” disposition, denoted by a green bar..



COMPANY_RESPONSE_TO_CONSUMER (event=Closed) Observations Used: **49,426**

COMPANY_RESPONSE_...

🟦 Closed
🟨 Closed with non-monetary relief
🟩 Closed with explanation
🟦 In progress
🟥 Closed with monetary relief
🟫 Untimely response

Name: Visualization 1

☑ Include missing

Frequency: Count

Growth strategy: Custom

Maximum branches: 3

Maximum levels: 6

Leaf size: 10

Response bins: 10

Predictor bins: 20

☐ Rapid growth

Pruning: Lenient Aggressive

☑ Reuse predictors

Assessment

☑ Use default number of bins

Number (5-100): 20

Prediction cutoff: 0.50

Tolerance: 0.000001

☐ Show diagnostic plots

☑ Auto-update

Figure 5: High-Level Depiction of COMPANY_RESPONSE_TO_CONSUMER Decision Tree

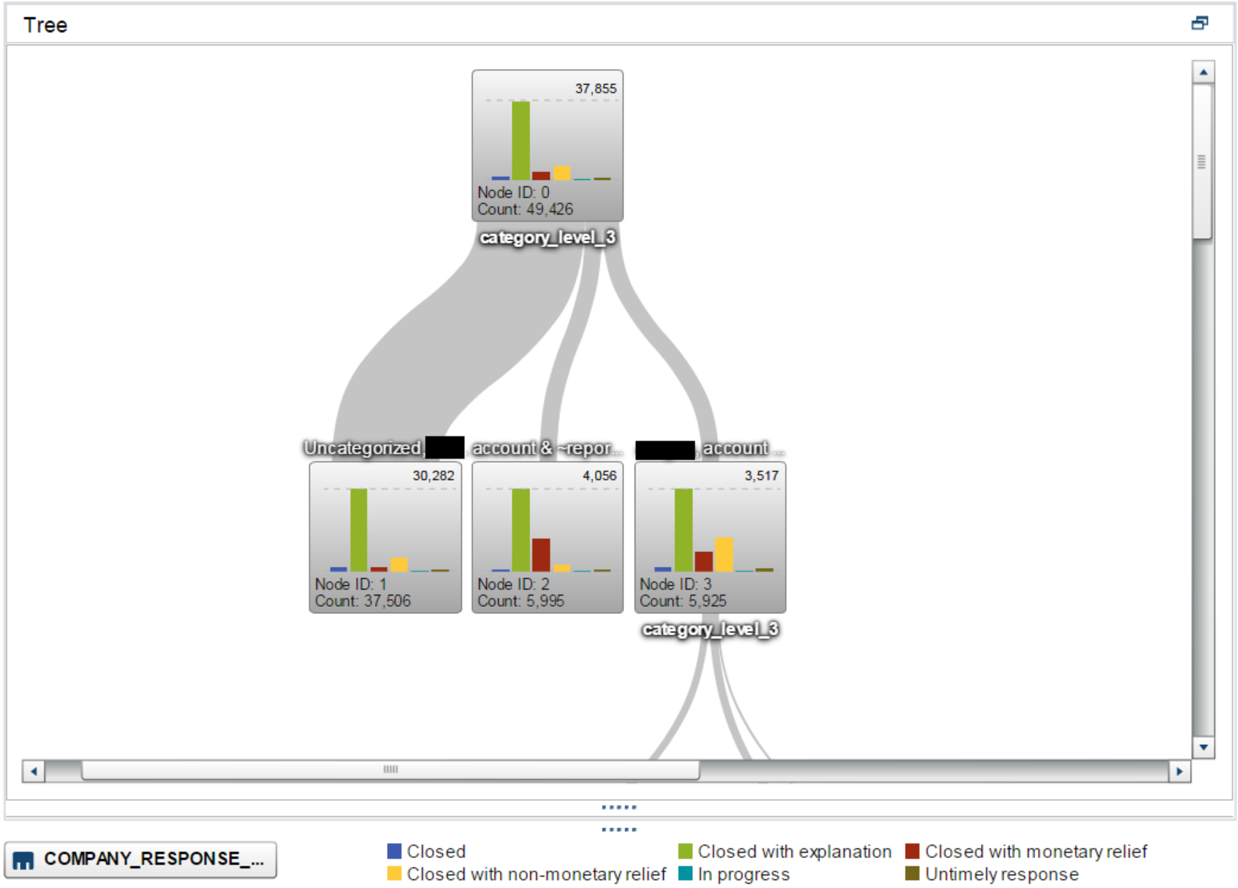
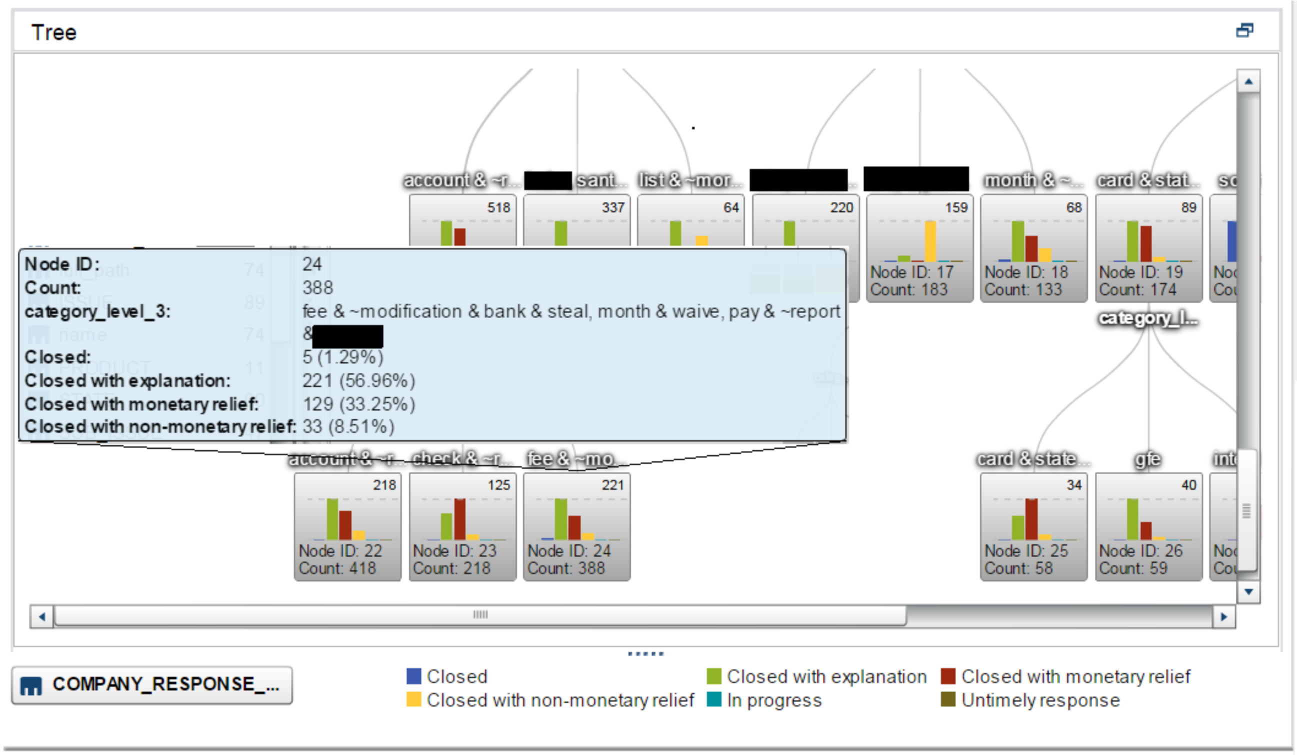


Figure 6: Top of the COMPANY_RESPONSE_TO_CONSUMER Decision Tree

Figure 7 shows how the proportion of responses that are closed with monetary relief dramatically increases as we traverse certain branches of the tree. There are three rules depicted in the highlighted node, including one that mentions the terms “fee”, “bank”, and “steal”, where “modification” is not mentioned. On the right hand side of the tree, note how one of the nodes is associated with “gfe” or good faith estimates, and that this warrants its own node that is strongly correlated with monetary relief. Insight garnered from this step is useful in explorations using the interactive reports.



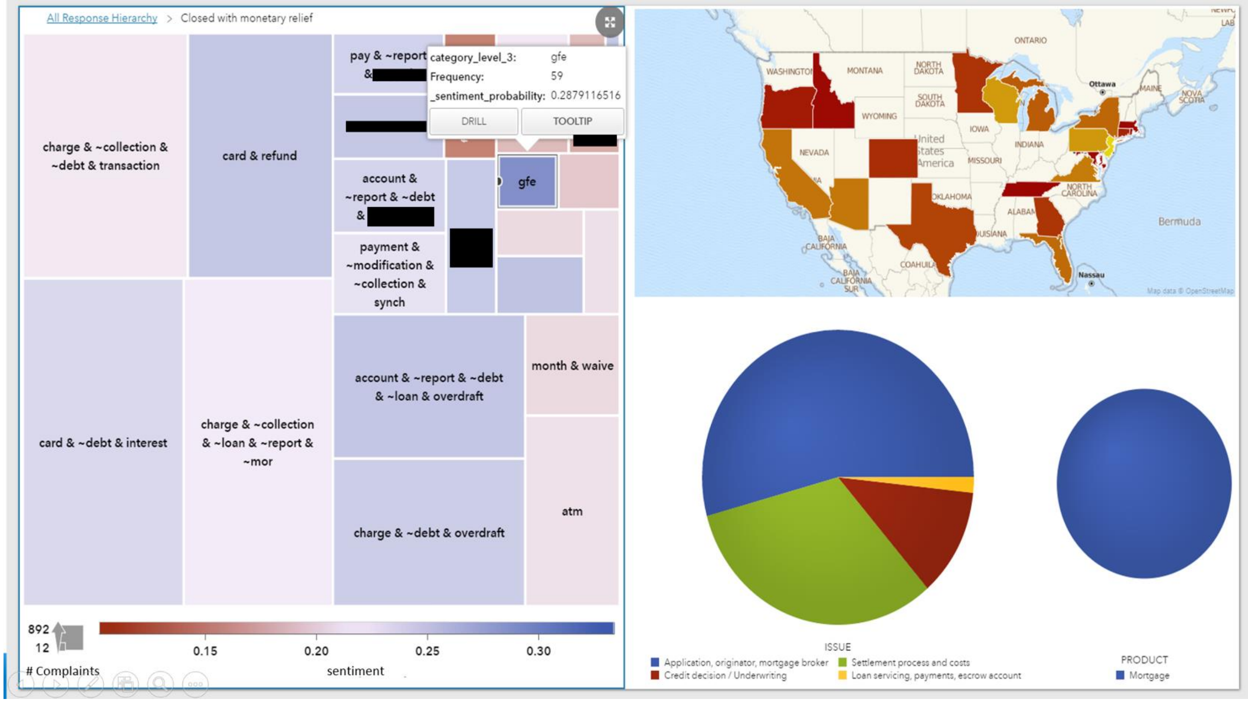
Node ID: 24
Count: 388
category_level_3: fee & ~modification & bank & steal, month & waive, pay & ~report & [redacted]
Closed: 5 (1.29%)
Closed with explanation: 221 (56.96%)
Closed with monetary relief: 129 (33.25%)
Closed with non-monetary relief: 33 (8.51%)

Figure 7: View of Decision Tree Highlighting Cases Strongly Associated with Monetary Relief

INTERACTIVE REPORT GENERATION AND USE

Interactive reporting enables the end-user analyst to explore the pre-existing data for complaints enhanced with the sentiment and rules that are generated from text analytics. This allows the analyst to sub-divide and prioritize exploration avenues according to the auto-categorization, while being guided by the relative levels of sentiment toward each of the categories. The analyst uses a dashboard, which depicts the rules and sentiment information in a tree map, the geospatial information and sentiment in a geospatial map, and information surrounding structured data issues and products in a pie chart. Links are provided from the tree and tile maps to drill down into the textual complaint data in a separate drilldown report. This drilldown report also includes a time series line chart so that analysts can observe trends over time.

Figure 8 depicts the use of a dashboard to explore one of the rules, “gfe” (good faith estimate). We have already determined in both SAS Contextual Analysis and SAS Visual Statistics that complaints containing this term are strongly correlated with monetary relief actions. From the dashboard, we can assess positive and negative sentiment, particularly at the geospatial level, so that we can begin to evaluate US states that might have been more particularly affected by misuse of the good faith estimate. We can also determine that this rule, as would be expected, entirely relates to mortgage products, and is associated with various issues surrounding the mortgage process.



category_level_3: gfe
Frequency: 59
_sentiment_probability: 0.2879116516

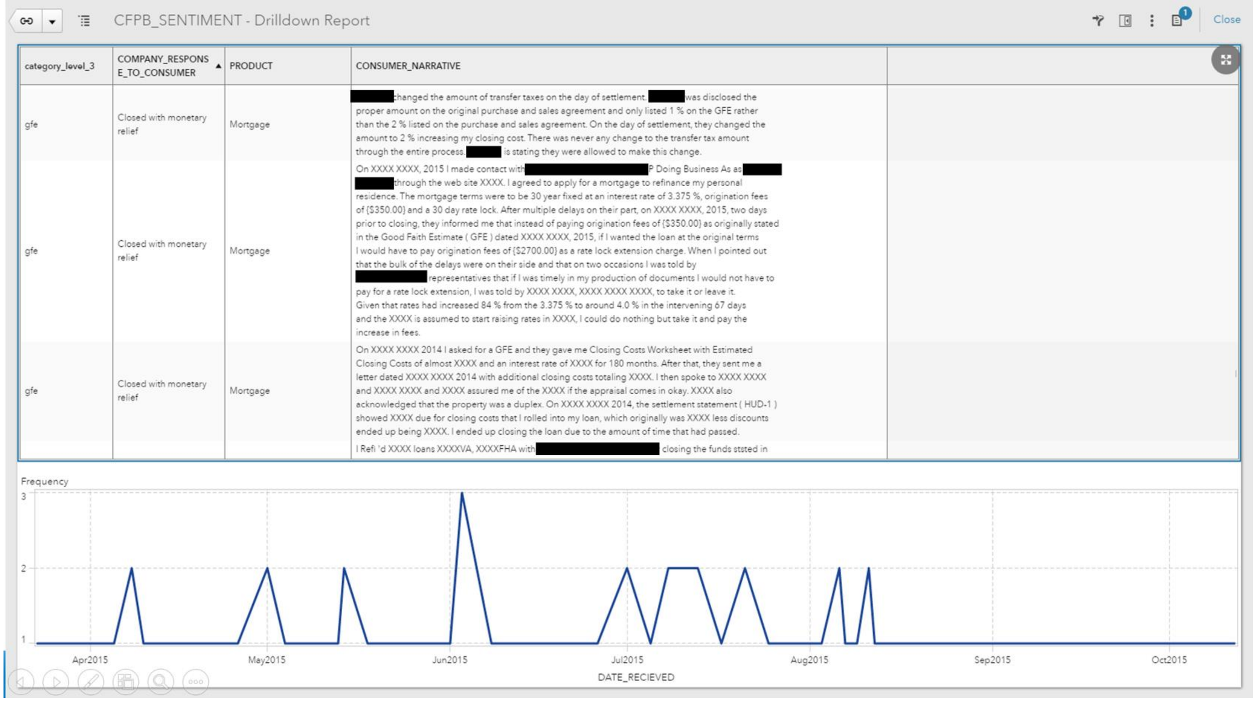
DRILL TOOLTIP

ISSUE
- 🟦 Application, originator, mortgage broker
- 🟩 Settlement process and costs
- 🟥 Credit decision / Underwriting
- 🟨 Loan servicing, payments, escrow account

PRODUCT
- 🟦 Mortgage

Figure 8: Dashboard Depicting Information Surrounding Good Faith Estimates in Complaint Data

Drilling down into the complaint data, as shown in Figure 9, we can see from the timeline that there are no new complaints related to good faith estimates after August 2015. This is because Congress directed the CFPB to combine two forms, the “Initial TIL Disclosure” and the “Good Faith Estimate” into the new “Loan Estimate and Closing Disclosure” form. This change took effect on October 3, 2015[6]. This new form is intended to be more transparent to the consumer, and therefore more difficult for financial organizations to misuse. The question remains, if analytics had been part of the assessment, could this ultimate decision to protect the consumer have been reached sooner?



category_level_3	COMPANY_RESPONSE_TO_CONSUMER ▲	PRODUCT	CONSUMER_NARRATIVE
gfe	Closed with monetary relief	Mortgage	[redacted]changed the amount of transfer taxes on the day of settlement. [redacted]was disclosed the proper amount on the original purchase and sales agreement and only listed 1 % on the GFE rather than the 2 % listed on the purchase and sales agreement. On the day of settlement, they changed the amount to 2 % increasing my closing cost. There was never any change to the transfer tax amount through the entire process. [redacted] is stating they were allowed to make this change.
gfe	Closed with monetary relief	Mortgage	On XXXX XXXX, 2015 I made contact with[redacted]P Doing Business As as[redacted]through the web site XXXX. I agreed to apply for a mortgage to refinance my personal residence. The mortgage terms were to be 30 year fixed at an interest rate of 3.375 %, origination fees of {$350.00} and a 30 day rate lock. After multiple delays on their part, on XXXX XXXX, 2015, two days prior to closing, they informed me that instead of paying origination fees of {$350.00} as originally stated in the Good Faith Estimate (GFE) dated XXXX XXXX, 2015, if I wanted the loan at the original terms I would have to pay origination fees of {$2700.00} as a rate lock extension charge. When I pointed out that the bulk of the delays were on their side and that on two occasions I was told by [redacted]representatives that if I was timely in my production of documents I would not have to pay for a rate lock extension, I was told by XXXX XXXX, XXXX XXXX XXXX, to take it or leave it. Given that rates had increased 84 % from the 3.375 % to around 4.0 % in the intervening 67 days and the XXXX is assumed to start raising rates in XXXX, I could do nothing but take it and pay the increase in fees.
gfe	Closed with monetary relief	Mortgage	On XXXX XXXX 2014 I asked for a GFE and they gave me Closing Costs Worksheet with Estimated Closing Costs of almost XXXX and an interest rate of XXXX for 180 months. After that, they sent me a letter dated XXXX XXXX 2014 with additional closing costs totaling XXXX. I then spoke to XXXX XXXX and XXXX XXXX and XXXX assured me of the XXXX if the appraisal comes in okay. XXXX also acknowledged that the property was a duplex. On XXXX XXXX 2014, the settlement statement (HUD-1) showed XXXX due for closing costs that I rolled into my loan, which originally was XXXX less discounts ended up being XXXX. I ended up closing the loan due to the amount of time that had passed.
			I Refi 'd XXXX loans XXXXVA, XXXXFHA with[redacted] closing the funds ststed in

Figure 9: Drilldown Report on Good Faith Estimate Complaint Narratives Including Timeline

One rule mentioned previously was indicative of “fees,” “stealing,” and “banks,” where the term modification is not mentioned. A dashboard depicting this rule is shown in Figure 10. By using this dashboard to explore the related complaints, we discovered that these complaints are split into two groups. One group includes cases where individuals claim that their banks are stealing from them based on the various fees these banks assess. The other group covers cases of identity theft and associated fees. This provides excellent feedback into the auto-generated taxonomy. SAS Contextual Analysis users can take this information and refine the rules for this particular sub-category to distinguish the identity theft cases from the excessive fees cases. After rescoring, they can better explore trends associated with the two new sub-categories. This example illustrates how statistical analysis goes hand-in-hand with capabilities to leverage subject matter expertise, refine the rules-based taxonomy, and better enable search and discovery.

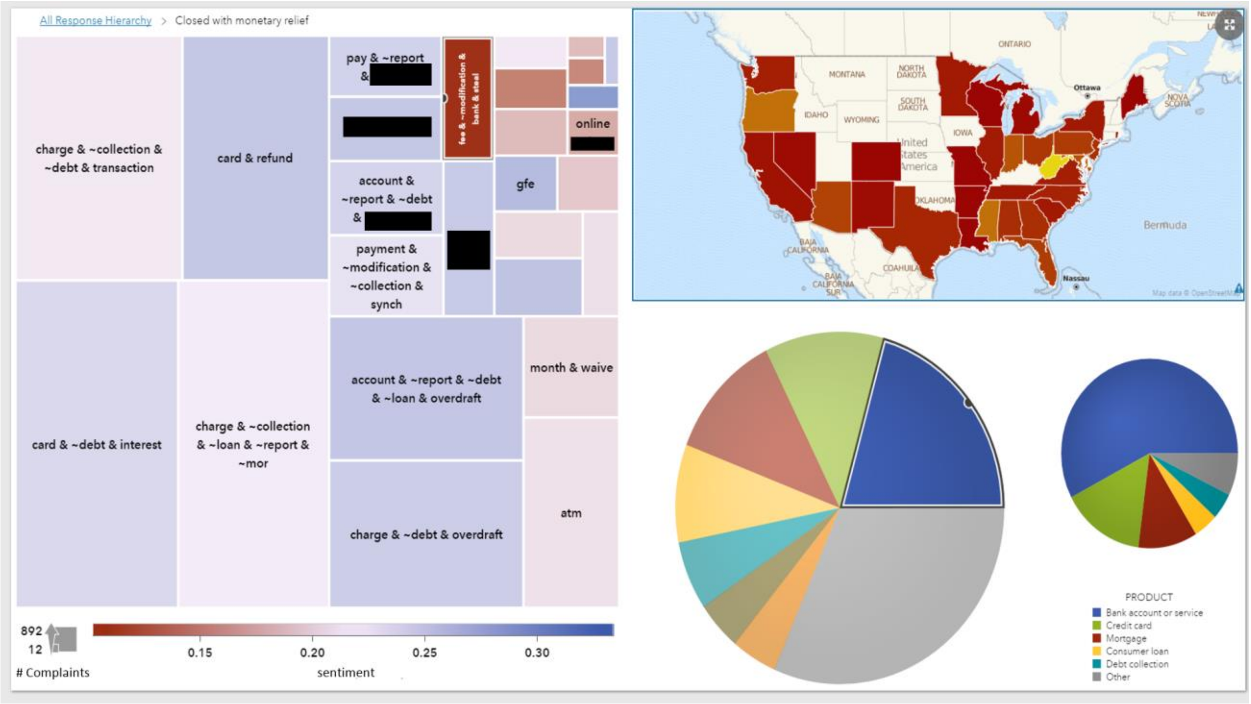


Figure 10: Dashboard Depicting Information Surrounding Excessive Fees and Identify Theft

CONCLUSION

In summary, we showcased a repeatable process that combined the benefits of both statistical and classification-based text analytics against the Consumer Financial Protection Bureau complaint data in order to assess these complaints for areas that trended toward monetary relief. In doing so, we identified several patterns, including one pattern that highlighted flaws related to good faith estimates, which is a part of the mortgage loan process on which CFPB has taken action. The quantitative analysis presented in this paper serves to validate the actions of the CFPB.

Overseeing organizations can use the methodology presented in this paper to improve time to value and quality of analysis when assessing complaint data. Financial organizations who support retail organizations through, for example, a debit card should use this methodology to help assess the quality of their customer care and their organizational satisfaction. Retail organizations should pay attention to assess whether the financial organization that is supporting their company is negatively impacting their brand.

The methodology depicted here is widely applicable. It relies on having substantial rows of data, generally 500 or more, in the context of a target variable of concern related to the text. The length of the text should typically be between a single line and several pages in order for this methodology to produce actionable information. An additional area where we can apply these capabilities is generating taxonomy around stand-up clinics after natural disasters. By analyzing a subset of the medical issues for which individuals are seen at these clinics, and for which we have a diagnosis code that differentiates between issues such as respiratory issues and bodily injuries, we can generate a taxonomy that characterizes these various issues in more detail. We can identify actionable information such as the type and quantity of materials that are needed at these clinics in order to ensure that medical needs are met for disaster survivors. For more areas of text analytics and subsequent visualization application, please see the SAS Global Forum paper, “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data[7].

The methodology also stands up to manual coding of textual data. Organizations that leverage machine learning capabilities can label social media content, for example, with tags that differentiate anything that needs to be analyzed. For example, an analyst could tag 1,000 Twitter entries related to food poisoning with a flag that differentiates actual instances of food poisoning. This can help build a model that more accurately identifies these instances from more generic talk. In an implementation of this example, the analyst might discover that certain terminology tends to surround the actual instances of food poisoning, such as the mention of a time-related term such as “hour(s)” or “day(s)”. Because this model also characterizes the tangible instances of food poisoning, an analyst might be interested in exploring all the cases in which the term “hour(s)” is mentioned because these might be more immediate. This is important if the analyst is looking for indications of a suspected epidemic.

A semi-automated feedback loop would enhance a machine-learning solution. In the context of this paper, this feedback loop is self-contained in SAS Contextual Analytics, enabling the user to modify the auto-generated categorical rules or to provide new complaint information in the context of the existing auto-generated rules. Feedback occurs when the user subsequently re-runs the models. Extension of this capability is something that should be considered, and could assist in determining whether documents end up fitting well in particular categorical buckets. Hence, a user of a visual exploration system would be able to dynamically re-assign primary reasons for monetary relief, and these re-assignments will be taken into account by the modeling software the next time the models are run.

The taxonomy, which was auto-generated and modified with subject matter expertise, could be used for auto-coding of new complaint data. It might be helpful for a reviewer to see, for example, that a new complaint matched a historical pattern such as issues with card refunds. It might also be helpful to see recommendations and contextual information surrounding the complaint. An example is “Here are a number of additional recent complaints matching the general pattern for the current one, related to card refunds, and here also are the general disposition for these complaints, such as how often they resulted in some form of monetary relief.” All of this information would assist in both the speed and quality of processing new complaints. This would not be difficult to implement using the SAS capabilities presented in this paper.

Finally, many data sources are not as structured as the data we obtained from the CFPB. For a demonstration of tokenization on a data set of large documents and subsequent analysis, see the SAS Global Forum papers “Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations”[8] and “Star Wars and the Art of Data Science: An Analytical Approach to Understanding Large Amounts of Unstructured Data”[9].

REFERENCES

1. Website of the Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/. Accessed on February 1, 2017.
2. Sabo, Tom. 2014. “Uncovering Trends in Research Using Text Analytics with Examples from Nanotechnology and Aerospace Engineering.” *Proceedings of the SAS Global Forum 2014 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS061-2014.pdf.
3. Sabo, Tom. 2015. “Show Me the Money! Text Analytics for Decision-Making in Government Spending.” *Proceedings of the SAS Global Forum 2015 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1661-2015.pdf.
4. Sabo, Tom. 2016. “Extending the Armed Conflict Location and Event Data Project with SAS® Text Analytics.” *Proceedings of the SAS Global Forum 2016 Conference*. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6380-2016.pdf.
5. “Consumer Complaint Database" Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data. Accessed on February 1, 2017.
6. “How we improved the disclosures” Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/know-before-you-owe/compare/. Accessed on February 1, 2017.

7. Sabo, Tom. 2014. SAS Institute white paper. “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data.” Available http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.

8. Albright, Russ. Cox, James. Jin, Ning. 2016. “Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations” *Proceedings of the SAS Global Forum 2016 Conference*. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6241-2016.pdf.

9. Osborne, Mary. Maness, Adam. 2014. “Star Wars and the Art of Data Science: An Analytical Approach to Understanding Large Amounts of Unstructured Data.” Proceedings of the SAS Global Forum 2014 Conference. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS286-2014.pdf.

ACKNOWLEDGMENTS

Thanks to Sophia Victor for providing insights into the visualizations that would best convey the text analytics results of this project to a wider audience. Thanks to Emily McRae, Sophia Victor, and Mary Beth Simmons for reviewing and refining this paper.

RECOMMENDED READING

- Chakraborty, G., M. Pagolu, S. Garla. 2013. *Text Mining and Analysis; Practical Methods, Examples, and Case Studies Using SAS®. SAS Institute Inc.*
- *Reamy, Tom. 2016. Deep Text; Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data*. Medford NJ: Information Today, Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Tom Sabo, Principal Solutions Architect
1530 Wilson Blvd.
Arlington, VA 22209
SAS Federal LLC
+1 (703) 310-5717
tom.sabo@sas.com
@mrTomSab
http://www.sas.com

SAS587-2017

Exploring the Art and Science of SAS® Text Analytics: Best practices in developing rule-based models

Murali Pagolu, Christina Engelhardt, and Cheyanne Baird, SAS Institute Inc.

ABSTRACT

Traditional analytical modeling, with roots in statistical techniques, works best on structured data. Structured data enables you to impose certain standards and formats in which to store the data values. For example, a variable indicating gas mileage in miles per gallon should always be a number (for example, 25). However, with unstructured data analysis, the free-form text no longer limits you to expressing this information in only one way (25 mpg, twenty-five mpg, and 25M/G). The nuances of language, context, and subjectivity of text make it more complex to fit generalized models. Although statistical methods using supervised learning prove efficient and effective in some cases, sometimes you need a different approach. These situations are when rule-based models with Natural Language Processing capabilities can add significant value. In what context would you choose a rule-based modeling versus a statistical approach? How do you assess the tradeoffs of choosing a rule-based modeling approach with higher interpretability versus a statistical model that is black-box in nature? How can we develop rule-based models that optimize model performance without compromising accuracy? How can we design, construct, and maintain a complex rule-based model? What is a data-driven approach to rule writing? What are the common pitfalls to avoid? In this paper, we discuss all these questions based on our experiences working with SAS® Contextual Analysis and SAS® Sentiment Analysis.

INTRODUCTION

In the last two decades, the world of analytics saw a lot of interest and research in analyzing data available in text. Extracting meaningful insights from text data is a Herculean task due to the fact that human language is complex, unstructured, nuanced and generally comes with a very low signal to noise ratio. It is a great advancement in science when humans can impart capabilities to machines for analyzing and interpreting text. SAS® Text Analytics depends on computationally intensive algorithms, statistical techniques and natural language processing methods. Broadly, there are two different methodologies that you can use for analyzing text data in the SAS world – the statistical approach and the linguistic approach (Chakraborty, Pagolu, and Garla, 2013).

In the statistical approach (also known as bag-of-words approach), the frequency of occurrence and co-occurrence of terms in the document collection (also known as the *corpus*) play a key role. Those numbers are generated in a table named the term-by-document matrix, and then condensed further by means of dimension reduction techniques such as singular value decomposition. In the linguistic approach, you deal with the semantics of the terms and the context in which they appear in each document, but not how frequently they appear across the entire corpus. You can develop linguistic rules that use keywords and phrases, Boolean operators, linguistic qualifiers, lemmatization (the capability to roll up term variations and inflections to the root word), part-of-speech recognition, regular expressions, and entity/fact definitions.

These two approaches fundamentally differ in the sense that the statistical approach characterizes an entire corpus by considering all the documents in the collection at once (*inter*-document analysis), whereas the linguistic approach only inspects a single document at a time, evaluating it in isolation against a set of predefined rules (*intra*-document analysis).

In the statistical approach, the large number of terms which make up the entire corpus are analyzed in order to discover topics or themes which depict the document collection. SAS® Text Miner calculates frequency weights (local weights) and term weights (global weights) based on factors such as document-specific term frequency, frequency of most frequent term, number of documents and the number of documents in which a term appear (Chakraborty, Pagolu, and Garla, 2013). While frequency weights help determine the importance of a term in the overall composition of a document, term weights help you understand which terms can better discriminate between the documents. This fundamental assumption that those terms that are moderately frequent across the corpus but are highly frequent within those documents in which they appear can very well discriminate between groups of documents is the basis for the unsupervised techniques such as clustering and text topic extraction in SAS® Text Miner. A text analytics model built using the statistical approach can also be termed probabilistic since it quantifies the probability of a document belonging to a particular cluster and then finds terms that can explain those clusters using the calculated weights.

By way of contrast, in the linguistic approach there is no such weighting mechanism prescribed. A rule-based model doesn't characterize the entire corpus in any way, rather every linguistic rule/definition in the model is evaluated for each document individually. It either extracts a piece of information from the full text, or classifies the document into zero, one, or multiple categories. A model developed using the linguistic approach is deterministic in nature, not probabilistic. Either a document satisfies a given rule completely or it doesn't; there is no ambiguity about the outcome.

SAS® Contextual Analysis offers its users the ability to develop linguistic rules to define a framework for classifying a corpus into pre-defined labels that are also known as categories. Developing linguistic rules requires using subject-matter expertise as well as an understanding of the grammatical structures and the nuances of the language in the corpus. SAS Contextual Analysis also offers the flexibility for the analyst to write rule-based models for extracting important pieces of information from each document, regardless of the statistical importance of those terms or phrases in the corpus. For certain types of data, the linguistic approach can yield higher accuracy than the statistical approach, although the tradeoff is that a linguistic model typically takes a longer period of time and more planning to develop.

As the field of text analytics matures and incorporates deep learning and dependency parsing, these two worlds tend to converge; the mathematical approach derives semantic and syntactic relationships between terms that can further inform linguistic rules. In this paper, our focus is on discussing best practices for textual data analysis using the linguistic approach and occasionally comparing the pros and cons with the statistical approach. Readers are expected to have a basic understanding of Boolean operators, qualifiers, regular expressions, concepts and fact extraction rules. We strongly encourage you to refer to *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*.

ASPECTS OF TEXTUAL DATA ANALYSIS USING LINGUISTIC APPROACH

Before we delve a little further into the intricacies of linguistic approach, let us understand what SAS offers in this space today. **Contextual Extraction** is the ability to extract important bits and pieces of information from documents using various types of rule definitions. For example, you can extract names of individuals holding office in important positions at the White House mentioned in news articles for the past year. This task might require a combination of knowledge on the grammatical structure or patterns that are predominant in that collection of articles, as well as a list of names of those holding key positions in the White House. In other words, both an understanding of how to model language *and* subject matter expertise are important. SAS® Contextual Analysis offers regular expressions, part-of-speech tags, and predicate logic that can be used to build simple to advanced contextual extraction definitions.

Document Classification is the term used for the process of assigning documents in the corpus to one or more categories as the use case demands. Depending on the use case, a document can be classified to one and only one category or can be assigned to more than one category. For a vast majority of use cases, a pre-defined set of categories is determined beforehand. The categorization model development framework (also known as *categorization taxonomy*) in SAS Contextual Analysis is built to largely take advantage of linguistic rule-based model development while also capitalizing on some aspects of the statistical approach as a precursor to the model development exercise. This hybrid approach works well for those who want the tool to produce a "starter" rule-based model which they can then manually modify to develop their final taxonomy. The best thing about this approach is that it doesn't require adequate amounts of training data with a defined target variable indicating the actual category to which a document belongs. In addition, there might be cases where certain documents don't belong to any of the pre-defined categories and you might miss them in your reporting if you never perform unsupervised learning to detect and capture these unexpected themes. It is a best practice to define an "Unknown" or

“Unclassified” category, either by defining appropriate linguistic rules to ensure the uncategorized documents fall in this bucket, or by labeling documents which do not match to any of your categories as “Unclassified” during model execution and post-processing.

Sentiment Analysis is the science of picking up the affinity of the author of the text towards a particular object or its features and classifying it into any of the following four categories considered widely as the industry standard: positive, negative, neutral, or unclassified. SAS Contextual Analysis offers automatic document-level sentiment, while SAS® Sentiment Analysis caters to the needs of advanced use cases that require a custom hierarchy for extracting sentiment towards specific objects and/or their corresponding features. Both contextual extraction and feature-level sentiment analysis lend themselves slightly better to the linguistic, rules-based approach.

APPROACHES TO RULE BUILDING: MACHINE VERSUS MANUAL VERSUS HYBRID

Assuming you have sufficient volumes of high quality training data, supervised machine learning models are quick to generate and are often more robust than a human-built taxonomy. That is, they often include term variants, term combinations, or topics that a person might not have intuitively come up with on his or her own. Different techniques are available for comparing machine learning models, and users can modify parameters and weighting strategies. The accuracy/speed-to-value ratio is very favorable, as is the scalability aspect.

However – although manual rule-writing is more time consuming than statistical text modeling, the linguistic approach more readily allows for advantages such as these:

- explicit incorporation of input from Subject Matter Experts (SMEs)
- good alignment with business objectives
- transparent and human-readable/editable logic (critical for highly regulated industries where models must be explainable)
- clean, business-interpretable outputs immediately ready for executive level reporting

As mentioned previously, the hybrid approach may be a good option. Let the machine do the bulk of the heavy lifting but then the analyst comes in afterward to add in subject matter expertise. Remove extraneous logic which may not be aligned with business goals.

MODEL VALIDATION

Having a good amount of hand-scored “gold-standard” data is ideal for effective model validation and testing the robustness of a model – you could train a statistical model on some of it, and use the remainder for validation. With data to use as ground truth, you can calculate precision and recall statistics, confusion matrices, and have benchmarks for percent accuracy. Unlike binary target variables in a typical predictive modeling situation with structured data elements, rule-based modeling for large scale taxonomies has typically several hundreds of categories. It means the confusion matrix is very complicated and difficult to digest. For the purpose of validating categories, we suggest that you generate as many gold-standard data sets as the number of individual categories you are building, with sufficient representative data to validate each category in the taxonomy. An additional benefit of having analysts and SMEs hand-scored data prior to modeling is that they internalize the keywords, patterns, and constructs that apply to that particular type of data. It gives a sense for what can and cannot be feasibly accomplished through text modeling. This knowledge will be invaluable as they build the rules and taxonomy.

MODEL TUNING

Realistic goals should be set for model accuracy (assuming gold-standard data from the previous section) taking these points into account:

- Data from sources such as social media posts or discussion forums will be noisier and harder to model accurately when compared with product reviews, news articles, claims notes, survey responses, and official documentation, etc. Informal chatter, conversation threads, types of slang and abbreviations used in the online vernacular are more dependent upon heuristics and context for interpretation than other sources. Set your expectations for accuracy according to your needs.

- It is important to assess how accurately the model needs to perform in order to meet your goals. For example: near-perfect accuracy might be required for high-impact use cases such as medical documents or legal contracts, but is less important in gauging directional market sentiment on products.
- Consider which is more important in your situation: sensitivity versus specificity? There can be a cost associated with a false positive case versus a false negative case. Developing a confusion matrix evaluating overall profit or risk based on the prediction accuracy can be useful in those cases.
- There will likely be dependencies within the category rules. For example, Category A is referenced in the rule definition for Category B. In such cases, tuning rule definition for Category A can affect the performance of Category B and both should be evaluated during testing.

PERFORMANCE AND SCALABILITY

When performing contextual extraction of entities, facts, or sentiment, SAS offers many options for rule types, some of which can do the jobs of others. It is a best practice to use the least powerful rule type that will accomplish the task at hand. In other words, don't use a sledge hammer for a nail!

Why? Because more advanced rule types come at a cost of computational intensity, which may affect model performance at execution. For example, don't use a REGEX rule to match a literal string (keyword/phrase), when a simple CLASSIFIER rule will achieve the same goal. Similarly, do not use a PREDICATE_RULE when a C_CONCEPT or a CONCEPT_RULE will do. This guideline is particularly relevant in situations where nested rule definitions are created, or concepts are referenced in other definitions. For example, consider the two PREDICATE_RULE concepts "RuleA" and "RuleB" below, which contain references to "ConceptA" through "ConceptE" which can be CLASSIFIER, CONCEPT, or REGEX based concept definitions.

```
RuleA:
PREDICATE_RULE:(x,y):(AND,(DIST_4,"_x{ConceptA}","_y{ConceptB}"),(NOT,"RuleB"
))
```

```
RuleB:
```

```
PREDICATE_RULE:(x):(AND ,"_x{ConceptC}","ConceptD",(NOT,"ConceptE"))
```

Concept "RuleA" is a PREDICATE_RULE which extracts matches when ConceptA and ConceptB are found within a distance of four words from each other only when there are no matches found anywhere in the document for the concept "RuleB". In "RuleB", the definition verifies the pass/fail conditions of its referenced concepts, yet nothing is done anywhere with the matched "ConceptC" string returned by "RuleB". Reference to the concept "RuleB" within the PREDICATE_RULE in "RuleA" works as a binary test of yes/no as to whether "RuleB" matches. In "RuleB" only one argument is returned, predicate logic is not necessary and we can convert the PREDICATE_RULE in "RuleB" to the following:

```
RuleB - CONCEPT_RULE:(AND,"_c{ConceptC}","ConceptD",(NOT,"ConceptE"))
```

Some additional best practices:

- Avoid building contextual extraction models when a categorization model is sufficient, as there are performance implications.
- Exercise caution using REGEX (regular expression) rules as they are powerful, yet computationally intensive. Unless written carefully to match patterns precisely, they can cause a lot of performance issues.
- When deciding whether to create one model for all sources versus a model per source, consider performance and whether it's worth running a lot of data through a model that you know won't fit it.
- Consider pre-processing data for performance improvements. For example, stripping out things like

html tags might reduce the processing time, even though having them in there doesn't affect your model. Similarly, consider removing useless tables or plain text representations of images or chart objects, and so on.

MAINTAINABILITY

If multiple users or groups will be building rule-based text models on shared data or for similar purposes, we encourage you to build and maintain centralized, reusable helper/intermediate entities, so that everyone has some base definitions to start with are consistent. Create base definitions/assets for concepts such as brand names, product names, synonyms, and misspelling lists.

A common question is: How can I evaluate the efficacy of specific rule parts, or sub-rules, within a single category/concept definition? This is possible only through regression testing where you need to start with one or two rules within a category which yield maximum recall and reasonable precision. Then as you keep adding more and more rules, you will see that while precision keeps climbing up, recall starts declining. These additional rules can be negation rules to exclude false positives. A well-developed classification model always strikes the best possible balance between precision and recall values. Figure 1 is an illustration of how adding linguistic rules to a category increases recall and precision initially, and then curves down with falling recall and rising precision values. Points marked 1 through 6 in the plot represent the sub-rules within a category. In this example, it is ideal to stop after adding the first four rules.

If you have gold-standard data that's been human verified, you will have a baseline to see if the new model is performing better or worse. Also, you might run diagnostics on the percentage of overall corpus that the model is encompassing. For example, if the original model has hits on 90% of the taxonomy and it's been verified that the remaining 10% are of no value, you should watch this percentage. If it dips to 70% or 80%, it might be time to perform more unsupervised topic exploration on the unclassified data to see what emerging topics you may find that your model is not capturing.

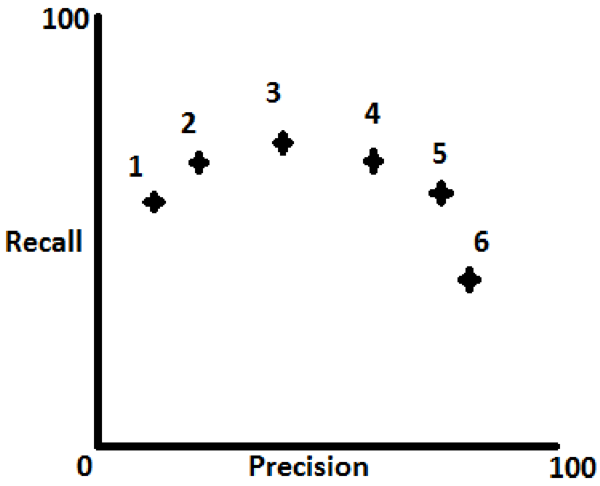


Figure 1. Recall versus Precision Plot for a Linguistic Rule-Based Category

ASSESSING THE IMPACT OF CATEGORY RULE MODIFICATION

It is quite common and feasible to periodically modify your category rules. In such instances, you might want to quantify how changing the rules affects your accuracy metrics (recall and precision) for that category. In the example below, Category1 is an existing rule comprised of certain keyword arguments, and a concept "Rule2" referenced as shown below:

Category1: `(OR, "term1", "term2", "term3", "term4")`

Category2 is similar to Category1 but has omitted terms term3 and term4.

Category2: `(OR, "term1", "term2")`

Now, a simple rule in Category3 like the one below can show documents that match only Category1 exclusively but not when they match Category2. Thus, this rule serves to provide an answer for you to understand how your rule modification has affected document matches and what documents you might have lost from dropping those two terms in Category2.

Category3:

```
(AND,
   _tmac:"@Top/Category1",
   (NOT,
      _tmac:"@Top/Category2"
   ))
```

In cases where you have added some terms, dropped some terms, or modified Boolean logic of any kind, it is a good idea to add another category as shown below while testing to know the impact in both directions, since you might have lost some matches and gained others.

Category4:

```
(AND,
   _tmac:"@Top/Category2",
   (NOT,
      _tmac:"@Top/Category1"
   ))
```

CAVEATS WHEN USING BOOLEAN RULE-BASED OPERATORS

With regard to the usage of Boolean linguistic rules in SAS Contextual Analysis, it is important to note a few things.

- Usage of the AND/NOT operator requires some diligence as they are essentially global in nature. This means that they are applied on the whole document and not limited by the boundaries of an encapsulating DIST or SENT operator. In other words, using AND/NOT operator within SENT or DIST operators will not yield the desired results; in these cases, you should use NOTINSENT or NOTINDIST operators.

  The NOTINSENT operator is used if you need to ensure a condition A is satisfied (that is, a certain set of terms need to appear) yet at the same time, you do not want another condition B to be found within the same sentence where condition A is satisfied. Please see the following examples to understand the usage of AND/NOT and NOTINSENT operators.

  Example 1 (AND/NOT):

  ```
  (AND,
     (OR, "term1", "term2", "term3"),
  (NOT,
     (OR, "term4", "term5", "term6")
  ))
  ```

  This rule will assign a document to this category if any of the desired terms 1, 2, or 3 are mentioned, but not if any of terms 4, 5, or 6 are present *anywhere in the entire document*. Wrapping the entire rule in an encompassing SENT or DIST operator will not change the global nature of this exclusion.

  Example 2 (NOTINSENT):

  ```
  (NOTINSENT,
     (OR, "term1", "term2", "term3"),
     (OR, "term4", "term5", "term6")
  )
  ```

This rule will assign a document to this category if any of the desired terms 1, 2, or 3 are mentioned, but not if any of terms 4, 5, or 6 are present *in the same sentence as the desired term*. Note that if terms 4, 5, or 6 occur elsewhere in the document, it will not prevent the match from occurring to this category. This logic is useful for removing unwanted contexts for a term at a local level, with the understanding that the excluded terms might be used in a valid way elsewhere in the document.

- In addition, use the following approach when you are writing sentence-level rules:

  If you need to verify the existence of conditions within a sentence, it is better to use the individual arguments for the SENT operator directly. The arguments for the SENT operator have an implicit AND relationship; nesting an explicit AND operator as well might not return the results you want. So, the rule below is not the correct way of using the SENT operator.

  Example 3 (Incorrect usage of SENT operator):

```
(SENT,
   (AND,
         (OR, "term1", "term2", "term3"),
         (OR, "term4", "term5", "term6")
   )
)
```

  Instead, use the following syntax.

  Example 4 (Correct usage of SENT operator):

```
(SENT,
   (OR, "term1", "term2", "term3"),
   (OR, "term4", "term5", "term6")

)
```

A DATA-DRIVEN APPROACH TO LINGUISTIC RULE-BASED MODELING

Rule-based model development can sometimes be a painstakingly long and exhaustive process. Depending on your ability to quickly discern patterns and how frequently they occur in the data, rules you develop manually might not effectively grab the majority of true positives and/or efficiently handle false positives. Any additional help in developing these rules can accelerate your rule development process. In this section, we will describe an innovative approach that assists in deriving the rules for categorization taxonomy.

Let us consider a use case where we are required to evaluate medical claims notes and assess which claims belong to the high risk category and which are at low risk. To demonstrate this approach, we created our own examples of sample claims notes where “smoking” is the morbid condition we are looking for in the claim adjustor notes. Our objective is to categorize a claim as high risk if we find at least one instance where it is mentioned in the notes by developing category rules that can detect true positive instances and exclude false positives. Now, in a realistic situation we should have some historical claims that we have manually evaluated as high risk versus low risk for such morbid conditions. Let us make an assumption that we have a “gold-standard”, hand-classified historical data set with the claims notes as well as an indicator telling us if those claims are high risk or low risk claims. Table 1 shows three examples highlighting portions of the text where smoking-related information is found in the claims notes.

Sample Text	Summary
Example 1: ……………………………… claimant has informed that he is **not a smoker** but drinks alcohol occasionally ……………he **smoked** for 10 years and then quit after he was diagnosed with ……………..	1 instance of True Positive, 1 instance of False Positive
Example 2: …………. ……………………………………………………………………………………….. ……………………………… alcohol: no, **smokes: yes** ………	1 instance of True Positive

Sample Text	Summary
Example 3: ……………………………………………………………he used to consume **2 packs per day** on an average which was below the usual for a regular ………………………………	1 instance of True Positive

Table 1: Sample Claims Notes with Instances of the Smoking Concept

Portions of text highlighted in blue indicate an instance of false positive context, while text highlighted in red indicates a true positive context. Our objective for this exercise is to capture all claims with at least one true positive context occurring anywhere in the claims notes.

A DATA-DRIVEN APPROACH METHODOLOGY

First, we will capture potential candidates for false positive cases by casting a wide net and catching the contexts for smoking-related terms wherever they occur in close proximity to negation terms. Using SAS Contextual Analysis, we can write a predicate rule with two extraction parameters, “neg” and “key”, to achieve this first step.

Example: `PREDICATE_RULE:(neg,key):(DIST_10,"_neg{NEGATIVE}","_key{SMOKING}")`

NEGATIVE – represents a Classifier concept for set of commonly occurring negation terms

SMOKING – represents a Classifier concept for SMOKING concept terms

neg – represents the parameter which captures the match returned by NEGATIVE concept from the text

key – represents the parameter which captures the match returned by SMOKING concept from the text

Here are some examples of terms that represent these concepts:

- Smoking – smoke, cigar, Chantix, tobacco, packs per day, nicotine etc.
- Negative – doesn’t, didn’t, denied, don’t, isn’t, ruled out, negative, no, none, non, false etc.

Using this predicate rule, we can extract the contexts from sample notes where a smoking concept keyword is within 10 tokens’ distance from a negative term. Once we extract the information from the sample notes, we can manually review the contexts to identify actual false positives and true positives. We might perform proximity analysis separately for hand-classified high risk sample claims notes simply based on any true positive contexts we might have found while analyzing the extracted contexts. By applying SAS Contextual Analysis scoring code, we can generate results with the “neg “and “key” parameters extracted with the help of the above PREDICATE_RULE. Using a SAS scan function, search the contexts for identifying the relative positions of the parameters (neg and key in this case). Table 2 shows some examples of how contexts are extracted from the sample claims notes. Proximity/distance between the keywords and relative direction are identified, and then separated as High risk and Low risk items. Occasionally, we may find several overlapping matches identified or extracted by a single PREDICATE_RULE if there are multiple keyword and negation term matches found in the document. In that case, we can consider the longest returned match for our analysis to get a better understanding of the overall context. When the relative positions of the extracted “neg” and “key” parameters are identified, we can record the proximity and direction as per these guidelines:

- **Proximity** – Distance between matches for “neg” and “key” parameters
- **Direction** – Position of negative term (neg) with respect to the concept keyword (key)

Note: + if neg occurs after key, - if neg occurs before key.

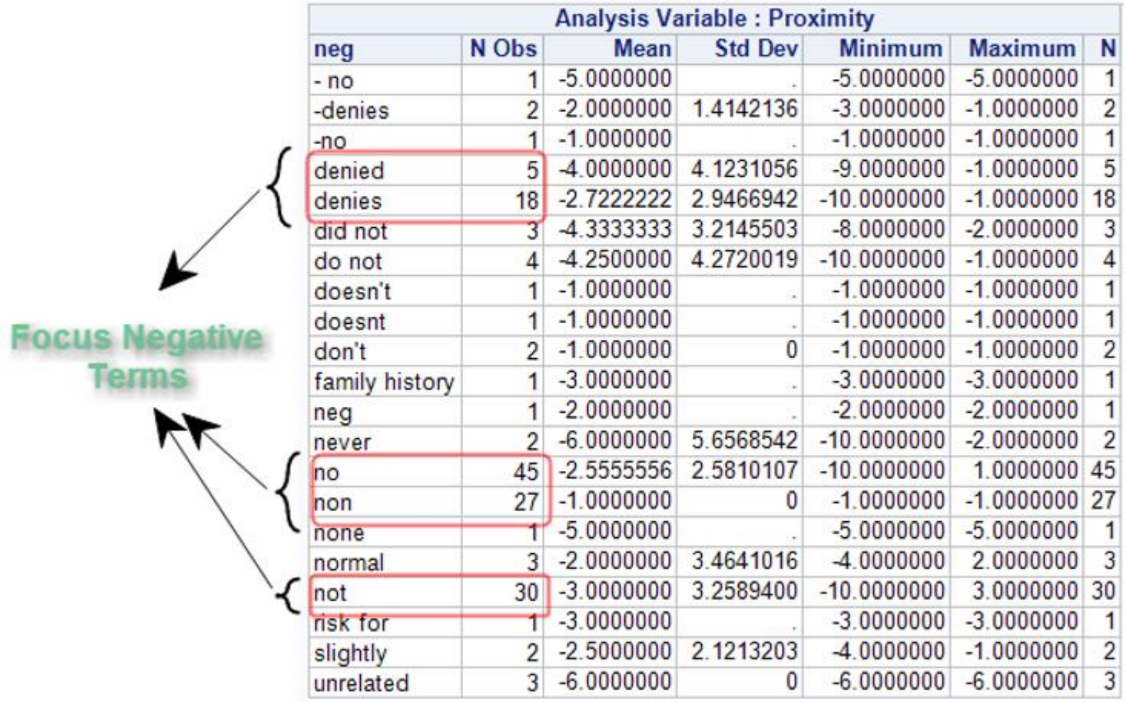
Classification (Smoking Concept)	Examples
Low Risk	……………not a smoker ……………[Distance: 2, Direction: -] ……………denies smoking …………[Distance: 1, Direction: -]

Classification (Smoking Concept)	Examples
	……………smoker : no ……………. [Distance: 2, Direction: +] ……………is not into smoking………. [Distance: 2, Direction: -] ……………is a non smoker ……….. [Distance: 1, Direction: -] ……………smokes - false ……………. [Distance: 2, Direction: +] ……………no tobacco ……………….. [Distance: 1, Direction: -]
High Risk	…… no alcohol but is a smoker….. [Distance: 5, Direction: -] ……….. alcohol: no, smokes: yes … [Distance: 2, Direction: -] ……didn't reveal he was a smoker…[Distance 5, Direction: -] …married no children clmt smokes…[Distance 3, Direction: -]

Table 2: Examples of False Positives and True Positives for the Smoking Concept

DISTRIBUTION ANALYSIS

As we derive the direction and distance/proximity values from the extracted information, we can study the distribution of the negation terms extracted as well as the average metrics for the proximity with the direction indicator applied. Display 1 shows a distribution analysis of negation terms with the summary statistics of proximity values on the sample set of claims notes. We can see that certain negation terms have high frequency over others, and they constitute a major portion of the cases with matches in the notes.


Focus Negative Terms

Analysis Variable : Proximity						
neg	N Obs	Mean	Std Dev	Minimum	Maximum	N
- no	1	-5.0000000	.	-5.0000000	-5.0000000	1
-denies	2	-2.0000000	1.4142136	-3.0000000	-1.0000000	2
-no	1	-1.0000000	.	-1.0000000	-1.0000000	1
denied	5	-4.0000000	4.1231056	-9.0000000	-1.0000000	5
denies	18	-2.7222222	2.9466942	-10.0000000	-1.0000000	18
did not	3	-4.3333333	3.2145503	-8.0000000	-2.0000000	3
do not	4	-4.2500000	4.2720019	-10.0000000	-1.0000000	4
doesn't	1	-1.0000000	.	-1.0000000	-1.0000000	1
doesnt	1	-1.0000000	.	-1.0000000	-1.0000000	1
don't	2	-1.0000000	0	-1.0000000	-1.0000000	2
family history	1	-3.0000000	.	-3.0000000	-3.0000000	1
neg	1	-2.0000000	.	-2.0000000	-2.0000000	1
never	2	-6.0000000	5.6568542	-10.0000000	-2.0000000	2
no	45	-2.5555556	2.5810107	-10.0000000	1.0000000	45
non	27	-1.0000000	0	-1.0000000	-1.0000000	27
none	1	-5.0000000	.	-5.0000000	-5.0000000	1
normal	3	-2.0000000	3.4641016	-4.0000000	2.0000000	3
not	30	-3.0000000	3.2589400	-10.0000000	3.0000000	30
risk for	1	-3.0000000	.	-3.0000000	-3.0000000	1
slightly	2	-2.5000000	2.1213203	-4.0000000	-1.0000000	2
unrelated	3	-6.0000000	0	-6.0000000	-6.0000000	3

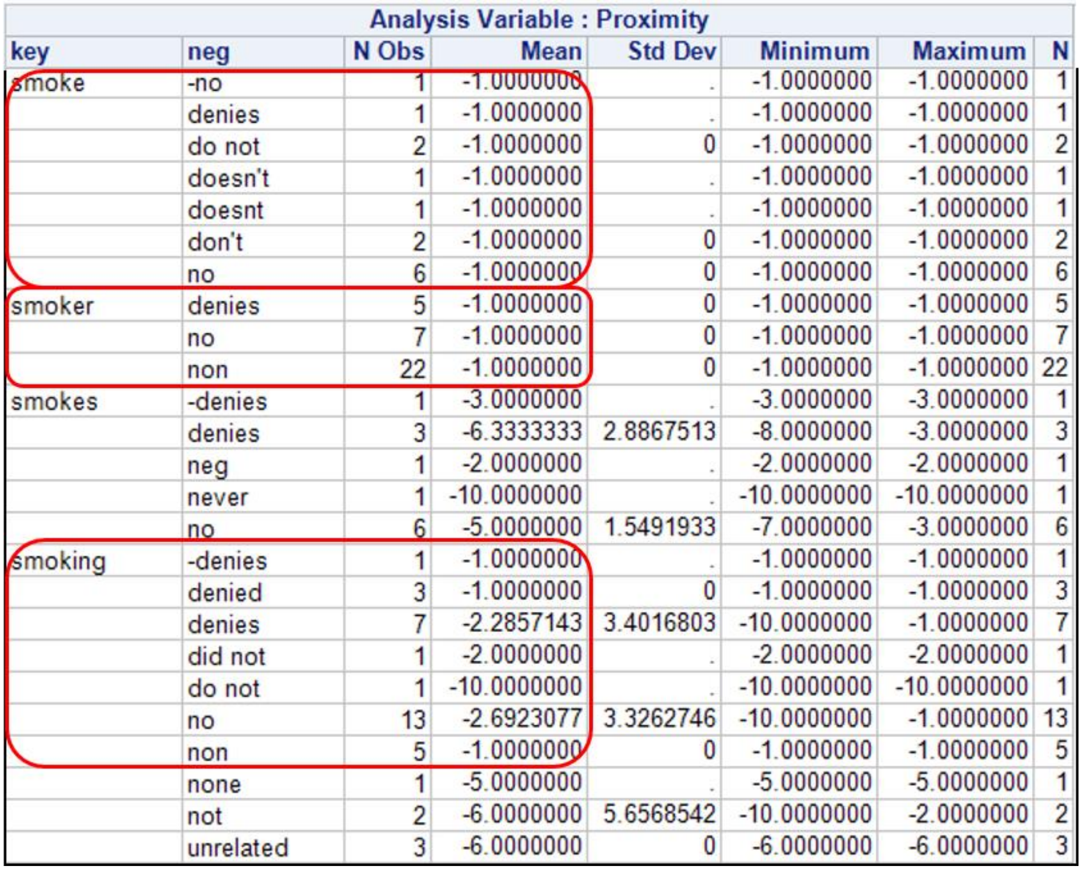
Display 1: Distribution of Frequency of Negation Terms

Along with the individual negation terms' frequency, we can also identify the most frequent combinations for negation and keywords for the smoking concept from the extracted information. Display 2 shows us the list of keyword and negation term combinations found in the sample notes along with the descriptive statistics of the proximity analysis variable. Again, looking at this distribution and how the mean proximity values show up, we can group certain negation terms and keywords in separate sets. We can write our own category rule as shown below, which qualifies as a data-driven rule based on our analysis and what story our data tells us.

Example: `(OR,(ORDDIST_5,”[SMOK_NEG_PRE]”,”[SMOK_1]”))`

SMOK_NEG_PRE – Negative terms that predominantly occur before the keyword concept terms.

SMOK_1 – Keyword concept terms that occur most frequently with the terms grouped under the SMOK_NEG_PRE classifier concept definition.



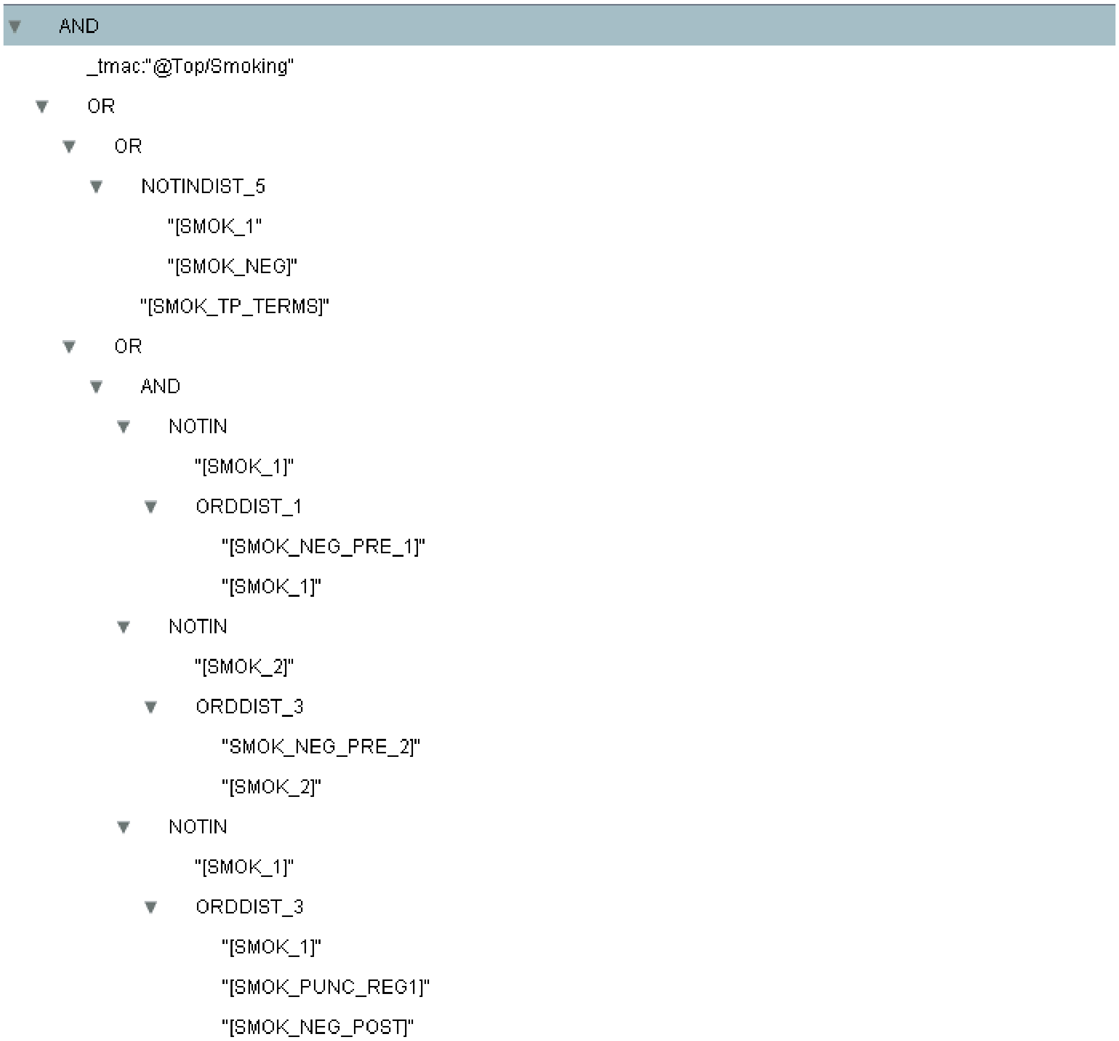
Analysis Variable : Proximity							
key	neg	N Obs	Mean	Std Dev	Minimum	Maximum	N
smoke	-no	1	-1.0000000	.	-1.0000000	-1.0000000	1
	denies	1	-1.0000000	.	-1.0000000	-1.0000000	1
	do not	2	-1.0000000	0	-1.0000000	-1.0000000	2
	doesn't	1	-1.0000000	.	-1.0000000	-1.0000000	1
	doesnt	1	-1.0000000	.	-1.0000000	-1.0000000	1
	don't	2	-1.0000000	0	-1.0000000	-1.0000000	2
	no	6	-1.0000000	0	-1.0000000	-1.0000000	6
smoker	denies	5	-1.0000000	0	-1.0000000	-1.0000000	5
	no	7	-1.0000000	0	-1.0000000	-1.0000000	7
	non	22	-1.0000000	0	-1.0000000	-1.0000000	22
smokes	-denies	1	-3.0000000	.	-3.0000000	-3.0000000	1
	denies	3	-6.3333333	2.8867513	-8.0000000	-3.0000000	3
	neg	1	-2.0000000	.	-2.0000000	-2.0000000	1
	never	1	-10.0000000	.	-10.0000000	-10.0000000	1
	no	6	-5.0000000	1.5491933	-7.0000000	-3.0000000	6
smoking	-denies	1	-1.0000000	.	-1.0000000	-1.0000000	1
	denied	3	-1.0000000	0	-1.0000000	-1.0000000	3
	denies	7	-2.2857143	3.4016803	-10.0000000	-1.0000000	7
	did not	1	-2.0000000	.	-2.0000000	-2.0000000	1
	do not	1	-10.0000000	.	-10.0000000	-10.0000000	1
	no	13	-2.6923077	3.3262746	-10.0000000	-1.0000000	13
	non	5	-1.0000000	0	-1.0000000	-1.0000000	5
	none	1	-5.0000000	.	-5.0000000	-5.0000000	1
	not	2	-6.0000000	5.6568542	-10.0000000	-2.0000000	2
	unrelated	3	-6.0000000	0	-6.0000000	-6.0000000	3

Display 2: Distribution of Frequency of Negation Terms and Smoking Concept Terms

RESULTS

Based on the above approach, we can develop and build several rules and then combine them under a single category. We create this category for identifying claims notes that have at least one instance of a true positive context for the smoking concept anywhere in the entire document. Display 3 is an example of how we can build a category rule based on the analysis we perform with negation terms and concept keywords on our data for a sample condition (smoking, in this case). In this example rule, we can see several rules stitched together using appropriate Boolean operators. This rule helps in catching claims notes with at least one true positive instance of the smoking concept mentioned anywhere in the document. This entire rule looks confusing and complicated, but as we develop these rules, make a habit of briefly noting down how we formulated these rules based on our analysis. In this way, maintainability and interpretability of these rules over a long period of time with or without our presence will be easier within our organization.

SMOK_1, SMOK_2, SMOK_NEG, SMOKE_NEG_PRE_1, SMOKE_NEG_PRE_2, SMOK_NEG_POST, SMOKE_TP_TERMS are all intermediate, or helper, CLASSIFIER concept definitions developed based on the distribution analysis of negation and keyword terms along with the proximity metric as described in the previous distribution analysis section. These helper concept definitions come in handy in building the individual rules that, when joined together, make one bulky category rule.



```
▼ AND
      _tmac:"@Top/Smoking"
  ▼ OR
      ▼ OR
          ▼ NOTINDIST_5
                "[SMOK_1"
                "[SMOK_NEG]"
            "[SMOK_TP_TERMS]"
      ▼ OR
          ▼ AND
              ▼ NOTIN
                    "[SMOK_1]"
                  ▼ ORDDIST_1
                        "[SMOK_NEG_PRE_1]"
                        "[SMOK_1]"
              ▼ NOTIN
                    "[SMOK_2]"
                  ▼ ORDDIST_3
                        "SMOK_NEG_PRE_2]"
                        "[SMOK_2]"
              ▼ NOTIN
                    "[SMOK_1]"
                  ▼ ORDDIST_3
                        "[SMOK_1]"
                        "[SMOK_PUNC_REG1]"
                        "[SMOK_NEG_POST]"
```

Display 3: An example showing a Category Rule for the Smoking Concept in SAS Contextual Analysis

NOTE: The default indention within the tree view (shown above) of the rule editor in SAS Contextual Analysis is really helpful for maintenance purposes when someone other than the developer of this rule needs to modify some portions of the rule. Imagine trying to initially comprehend the rule's logic if displayed in text view!

```
(AND,_tmac:"@Top/Smoking",(OR,(OR,(NOTINDIST_5,"[SMOK_1","[SMOK_NEG]"),"[SMOK
_TP_TERMS]"),(OR,(AND,(NOTIN,"[SMOK_1]",(ORDDIST_1,"[SMOK_NEG_PRE_1]","[SMOK_
1]")),(NOTIN,"[SMOK_2]",(ORDDIST_3,"SMOK_NEG_PRE_2]","[SMOK_2]")),(NOTIN,"[SM
OK_1]",(ORDDIST_3,"[SMOK_1]","[SMOK_PUNC_REG1]","[SMOK_NEG_POST]"))))))
```

In the following paragraph, we show you an example of how you can interpret and explain a complicated category rule such as this.

Categorize a claims-notes document as a "True Positive" for the 'Smoking' concept if,

a) A 'smoking' concept term is found anywhere in the notes

AND

b) At least one of the following two conditions listed in 1 and 2 is satisfied:

1. A ‘smoking’ related concept term from SMOK_1 classifier concept definition is found such that it is not within a distance of 5 words from a ‘negative’ term (SMOK_NEG)

   (Or)

   Any term strongly suggesting that the claimant is a smoker (SMOKING_TP_TERMS)

   Examples: long term smoker, significant smoker, Chantix, emphysema, varenicline, packs a day, packs per week, pks/day, and so on.

2. A ‘smoking’ related concept term from SMOK_1 is found such that **it is not** within an ordered distance of 1 word from a negative term found in SMOK_NEG_PRE_1 concept definition.

   Examples: non smoker, doesn’t smoke, no smoking, no tobacco, and so on.

   (And)

   A ‘smoking’ related concept term from SMOK_2 is found such that **it is not** within an ordered distance of 3 words from any of the negative terms found in SMOK_NEG_PRE_2 concept definition.

   Examples: not a smoker, not into smoking, no habit of smoking, denies use of tobacco, negative for smoking, and so on.

   (And)

   A ‘smoking’ related concept term from SMOK_2 is found such that **it is not** followed by a punctuation mark (: - \ /) and any of negative terms from SMOK_NEG_POST in that particular order.

   Examples: smoker – no, smokes ? No, smoker : false, smoking – neg, smoke – denied, etc.

Note: Regardless of any number of false positive contexts identified in a claims-notes document, this rule will override them and tag the entire notes as “true positive”.

Summary of Data-driven Approach

- Using data-driven analysis helps you to develop linguistic rule-based categorization or contextual extraction models based on the patterns found in the data.
- Contextual extraction rules help you to understand the patterns in the data.
- Using powerful factual extraction rules such as PREDICATE RULE, you can not only extract the matching parameters (concept keyword and negative term) but also extract the concordance (a certain number of words or characters before and after the matching context).
- Using both categorization and contextual extraction features simultaneously has its own benefits. However, exercise caution when using REGEX (regular expression) rules since you might be matching several thousand terms in the documents with a small mistake in the rule.
- Performing the distribution analysis of negation and keyword terms over proximity measure separately for hand-classified “gold-standard” High risk Versus Low risk documents will help develop precise rules.
- The categorization rules framework in SAS Contextual Analysis provides powerful operators to incorporate negation scenarios to exclude false positive contexts very easily.

GENERAL BEST PRACTICES FOR RULE-BASED SENTIMENT ANALYSIS MODELING

As when creating categorization or contextual extraction taxonomies, for sentiment analysis we also encourage you to create intermediate entities as “helpers”, or reusable building blocks, that you can reference in other rules. This allows you to create definitions to capture the essence of a generic feature such as “Price”, and then just combine that definition with other contextual indicators, such as a specific product or brand, in order to compare sentiment toward Price in a fair fashion across products and brands. In addition, by creating this symbolic reference to the single Price definition, it is simple to later add more keywords and rules in that single place to extend the definition and have those updates automatically propagate to the other rules that reference it. This makes for very efficient, interpretable, and maintainable models.

Keep in mind that some words are not universally positive or negative – the tone depends on the context. For example, the word “cheap” might be positive in the context of price, but negative in the context of product quality. In such cases, you can add these ambiguous keywords to the feature level positive/negative tonal keyword lists, rather than the global lists.

SAS® Sentiment Analysis Studio supports testing folders for gold-standard data that is pre-classified as positive, negative, and neutral. Whenever possible, we encourage you to upload gold-standard data in this format to simplify the process of testing and enable you to readily see what type I and type II errors your rules are generating.

It can be tempting to try to capture every single nuance of language correctly in your sentiment model right away. For example, you might want to properly detect the following:

- Sarcasm (for example, “Great. The train is late AGAIN!”)
- Conditional statements (for example, “If the service had been better we would have come back.”)
- Model operators (for example, “I might consider upgrading to the new model”, “The food could’ve been better.”)
- Comparatives and Amplifiers (for example, “it was way worse than before.”)
- Negation (for example, “I was not happy with the experience.”, “The crew could not have been nicer!”)

While these more subtle language patterns can sometimes be modeled with a bit more effort, we recommend that you handle the more straightforward expressions of sentiment first, and then tackle these more complex cases in subsequent phases of tuning. (Negation is the exception; this pattern can be captured and assigned properly in most cases with a few additional pieces of logic, and is typically part of an initial sentiment model.)

In situations where the above types of tricky cases only represent a small percentage of your overall corpus and the rules added to catch them carry the risk of losing good matches or causing false positives in the rest of your data, it may be worth ignoring these edge cases in your model altogether and chalking them up to misclassification error. Remember – even humans only agree on how to interpret text 80% of the time; do not expect your models to be perfect!

CONCLUSION

The statistical approach is fast and easy to maintain. It easily scales up to increasing data volumes or changing data patterns for exploratory and predictive modeling needs. The linguistic approach is a time-consuming and sophisticated process, but can yield incrementally more accurate results if the assets are built using domain knowledge and subject matter expertise and the models are well-maintained. In our experience, the statistical approach and the linguistic rule-based approach each have their own benefits and drawbacks. Depending on the use case or application purpose, one might take precedence over the other. Generally, one approach outperforms the other depending on the nature of data and objective/goal of the analysis. In our experience, the statistical approach works best for internal data such as surveys, call center logs, manufacturer warranty claims, technician notes, and so on, where exploration of the data for generating themes or predictive modeling is the priority. Linguistic rule-based modeling is best suited for applications requiring classification of documents into pre-determined categories/sub-categories and contextually extracting information from dense documents such as medical claims notes, legal documents, academic publications, and so on. In those cases, it is important to contextually verify the occurrence or absence of desired concepts to disambiguate between false positives versus true positives. Text Analytics is as much an art as it is a science, and each individual use case offers its own unique opportunity for you to apply creativity, data mining techniques, and domain knowledge to best solve the problem at hand.

REFERENCES

Aizawa, A. 2003. “An Information-Theoretic Perspective of tf-idf Measures.” *Information Processing & Management*. 39 (1): 45-65.

Booth, A. D. 1967. “A ‘Law’ of Occurrences for Words of Low Frequency.” *Information and Control.* 10 (4): 386-393.

Chakraborty, G., M. Pagolu, and S. Garla. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. Cary, NC: SAS Institute Inc.

Manning, C. D., and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Text Analytics Using SAS Text Miner. SAS Institute course notes. Course information: https://support.sas.com/edu/schedules.html?ctry=us&id=1224

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

ACKNOWLEDGMENTS

We would like to thank Saratendu Sethi, Praveen Lakkaraju, Teresa Jade, and Michael Wallis from SAS for their support and guidance over the years.

RECOMMENDED READING

- *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*

CONTACT INFORMATION

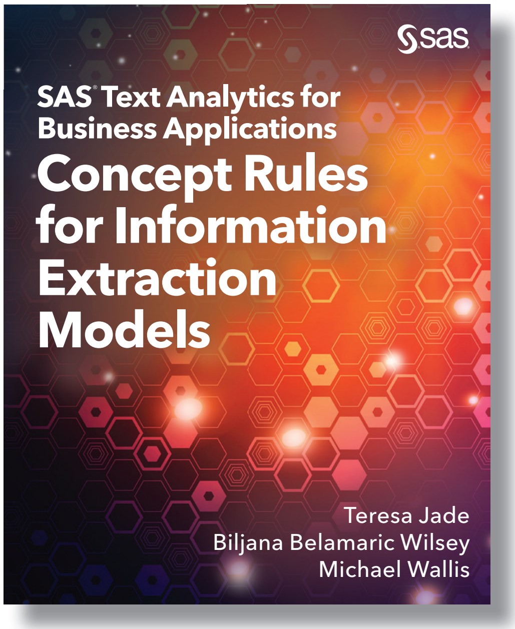
Your comments and questions are valued and encouraged. Contact the authors:

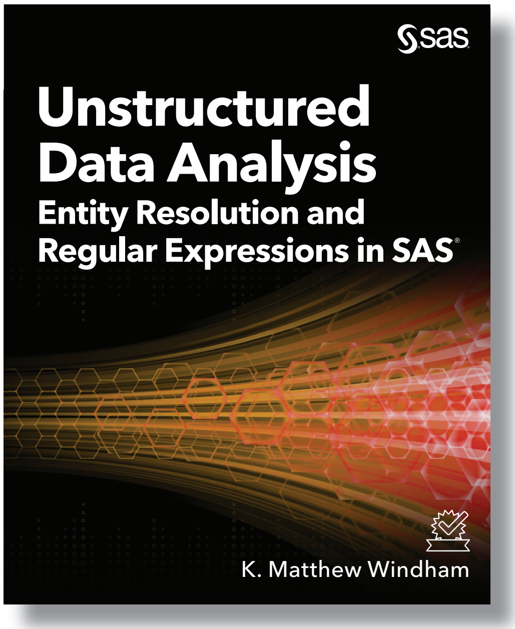
Murali Pagolu
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Murali.Pagolu@sas.com
http://www.sas.com

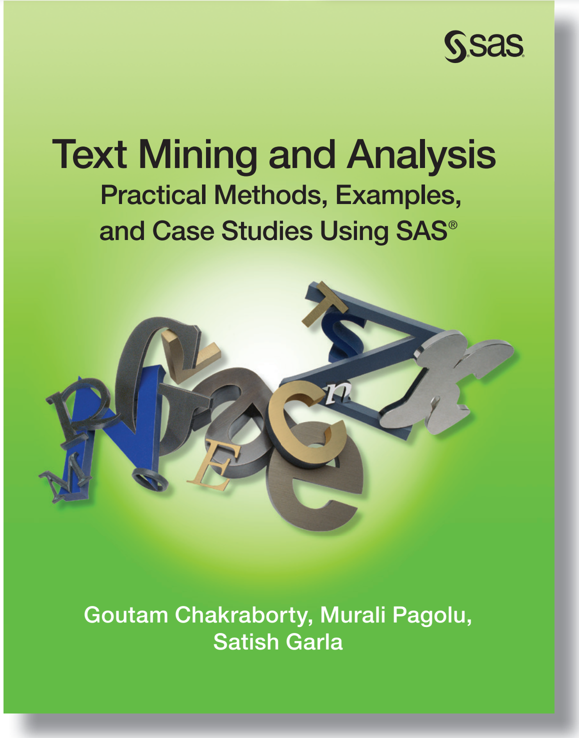
Christina Engelhardt
124 Preserve Way
Mooresville, NC 28117
SAS Institute Inc.
Christina.Engelhardt@sas.com
http://www.sas.com

Cheyanne Baird
SAS Institute Inc.
Cheyanne.Baird@sas.com
http://www.sas.com

For more information on this topic, check out the below books in the SAS® bookstore:


sas
SAS Text Analytics for Business Applications
Concept Rules for Information Extraction Models
Teresa Jade
Biljana Belamaric Wilsey
Michael Wallis


sas
Unstructured Data Analysis
Entity Resolution and Regular Expressions in SAS®
K. Matthew Windham


sas
Text Mining and Analysis
Practical Methods, Examples, and Case Studies Using SAS®
Goutam Chakraborty, Murali Pagolu, Satish Garla

For 20% off these e-books, visit **sas.com\books** and use **WITHSAS20**

sas.com/books
for additional books and resources.

SAS® THE POWER TO KNOW®